e-ISSN 2828-0490
JURNAL MEDIA COMPUTER SCIENCE

Multimodal Learning in AIoT Systems: Sensor Fusion and Vision-Based Intelligence

Agnes Prima Wulanjari [1)]; Ria Dymyati [2)]; Indar Bismoko [3)]; Nuryake Fajaryati [4)]; Pipit Utami [5)]

[1,2,3,4,5)] ***Department of Distance Education in Technical and Vocational Education, Universitas Negeri Yogyakarta***

Email: [1)] dedyirvandy.2024@student.uny.ac.id

How to Cite :

Wulanjari, A, P., Dymyati, R., Bismoko, I., Fajaryati, N., Utami, P. (2025). Multimodal Learning in AIoT Systems: Sensor Fusion and Vision-Based Intelligence. Jurnal Media Computer Science, 5(2)

ARTICLE HISTORY

Received [22 Juni 2025]
Revised [30 Juli 2025]
Accepted [31 Juli 2025]

KEYWORDS

Multimodal Learning; Artificial Intelligence of Things (AIoT); Sensor Fusion; Computer Vision; Meta-Analysis; Performance Evaluation.

ABSTRAK

Penelitian ini mengevaluasi efektivitas pembelajaran multimodal dalam sistem Artificial Intelligence of Things (AIoT), dengan fokus pada integrasi sensor fusion dan computer vision untuk tugas klasifikasi. Metode yang digunakan adalah systematic review dan meta-analisis terhadap studi yang dipublikasikan pada periode 2020–2025. Sebanyak 13 studi memenuhi kriteria inklusi, namun hanya 6 studi yang menyediakan data kuantitatif yang dapat dibandingkan akibat keterbatasan pelaporan baseline dan praktik evaluasi. Hasil menunjukkan bahwa pendekatan multimodal umumnya meningkatkan akurasi dibandingkan baseline unimodal ketika evaluasi yang sebanding tersedia, dengan rata-rata peningkatan sebesar 8,88% (95% CI: 5,33%–12,44%, p < 0,001). Heterogenitas yang tinggi ditemukan, dipengaruhi oleh domain, konfigurasi sensor, dan arsitektur model. Temuan ini menunjukkan bahwa efektivitas multimodal bersifat kondisional dan bergantung pada komplementaritas modalitas, strategi fusi, serta batasan sistem.

ABSTRACT

This study evaluates the effectiveness of multimodal learning in Artificial Intelligence of Things (AIoT) systems, focusing on the integration of sensor fusion and computer vision for classification tasks. A systematic review and meta-analysis were conducted on studies published between 2020 and 2025. Thirteen studies met the inclusion criteria; however, only six provided comparable quantitative data due to inconsistent baseline reporting and evaluation practices. The results indicate that multimodal approaches generally improve accuracy compared to unimodal baselines when comparable evaluations are available, with an average increase of 8.88% (95% CI: 5.33%–12.44%, p < 0.001). High heterogeneity was observed, influenced by domain, sensor configuration, and model architecture. These findings suggest that multimodal effectiveness is conditional and depends on modality complementarity, fusion strategy, and system-level constraints.

INTRODUCTION

Artificial Intelligence of Things (AIoT) is increasingly used in real-world systems that require continuous sensing and timely decision-making, such as healthcare monitoring, smart agriculture, and industrial inspection. In these settings, sensing is not just about collecting data, but about ensuring that the system can interpret its environment reliably under practical constraints. Limited

computational resources, latency requirements, and heterogeneous data sources make perception in AIoT systems a non-trivial problem that directly affects system performance.

Computer vision plays a major role in AIoT because it provides rich semantic information from visual inputs. However, relying on vision alone is often not sufficient in practice. Unimodal vision-based approaches are sensitive to environmental factors such as lighting changes, occlusion, and noise. They are also affected by domain shift and limited labeled data when deployed outside controlled settings. In addition, running deep vision models on edge devices introduces constraints in computation, energy consumption, and latency, which can reduce responsiveness and robustness in real deployments (Lin, 2023; J. Zhang & Tao, 2021). These issues suggest that vision-only approaches may struggle to maintain stable performance in dynamic AIoT environments.

To address these limitations, multimodal learning and sensor fusion have been explored as practical alternatives. By combining visual data with other modalities, such as physiological signals, environmental sensors, or contextual information, systems can form a more complete representation of the observed environment. This integration helps reduce the weaknesses of individual modalities and improves robustness under varying conditions. In addition, distributing computation across edge, fog, and cloud layers allows multimodal systems to balance latency, resource usage, and accuracy more effectively (Lei et al., 2020; Lin, 2023; D. Wu et al., 2024). In practical AIoT implementations, however, the integration of multiple sensing modalities does not automatically translate into effective system-level performance. Prior studies in IoT-based agricultural systems highlight that increasing sensing and analytical capabilities often lead to monitoring-oriented systems, where data collection is emphasized but its translation into operational decisions remains limited (Utami, Zakarijah, et al., 2024).

Despite these reported advantages, a critical gap remains in the literature. Existing studies predominantly emphasize performance improvements of multimodal approaches but rarely examine whether these gains are consistent, generalizable, or dependent on specific conditions. Reported improvements are often tied to particular datasets, model configurations, or experimental setups, while differences in evaluation protocols, the absence of standardized baselines, and inconsistent reporting practices hinder cross-study comparability (Ayetiran & Özgöbek, 2024; L. Wu et al., 2023; Z. Zhang et al., 2020). This limitation reflects a broader pattern observed in AIoT research, where technological components are frequently developed in isolation, resulting in fragmented system integration and limited operational impact (Utami, Mashoedah, et al., 2024). Such findings suggest that performance improvements should not be interpreted solely at the model level, but also in relation to system-level integration and deployment constraints. As a result, the field lacks a unified understanding of whether multimodal learning constitutes a generally superior paradigm or a context-dependent strategy whose effectiveness varies across system configurations.

This study addresses this gap by explicitly positioning multimodal learning as a conditional, rather than universally beneficial, approach. We argue that although multimodal learning generally improves performance over unimodal approaches, these gains are not inherent to modality integration itself. Instead, they emerge from the interaction between three key factors: modality complementarity, fusion strategy, and system-level constraints. This perspective shifts the focus from “how many modalities are used” to “how effectively modalities are selected, combined, and deployed within system constraints.”

Meta-analysis provides a systematic approach to move beyond isolated experimental results by aggregating quantitative evidence across studies and examining variability in effect sizes. This is particularly important in AI research, where experimental diversity often limits direct comparison. In this study, a meta-analysis is conducted on multimodal approaches in AIoT systems, focusing on classification tasks that integrate computer vision with additional sensor modalities. From the collected studies, only a subset provides comparable quantitative results with both unimodal and multimodal performance, enabling effect size estimation. The results indicate that multimodal approaches generally improve accuracy when comparable evaluations are available, while also

revealing substantial variability across studies, highlighting that performance gains are context-dependent rather than uniform.

This study makes three main scientific contributions. First, it provides quantitative evidence through meta-analysis that multimodal learning generally improves performance compared to unimodal approaches under comparable evaluation settings. Second, it explains the high variability in reported performance gains by systematically linking them to differences in domain, modality configuration, and model architecture, consistent with the heterogeneity observed in empirical results. Third, it develops a structured conceptual framework that integrates modality complementarity, fusion strategy, and system-level constraints as key determinants of multimodal effectiveness in AIoT systems, aligning quantitative findings with system-level interpretation.

The novelty of this study lies in reframing multimodal learning performance as a conditional phenomenon rather than a universally beneficial approach. Unlike prior studies that primarily report performance improvements in isolation, this work introduces a unified perspective that connects performance variability to fundamental design factors, supported by both quantitative synthesis and conceptual modeling. This reframing provides a more realistic and actionable understanding of multimodal AIoT systems, particularly in real-world deployment scenarios where constraints and variability are unavoidable. Together, these findings aim to clarify when and under what conditions multimodal approaches provide meaningful benefits in AIoT systems.

THEORETICAL BACKGROUND

The conceptual basis for this analysis is framed around how key variables are defined and related within AIoT-based perception systems, with attention to practical issues observed in the reviewed studies, such as inconsistent reporting, variation in evaluation practices, and the limited availability of comparable baselines. Table 1 summarizes the variables by explicitly presenting their roles, descriptions, and corresponding components, including how multimodal and unimodal approaches are compared, what outcome is measured, and which factors are considered to influence performance across studies. The studies analyzed in this work were collected from major academic databases, including Scopus, Wiley, and ScienceDirect.

Table 1. Conceptual Framework and Variables Used in The Analysis

Role	Description	Component
Independent Variable	Comparison between multimodal approaches (sensor fusion + vision) and unimodal baselines	Multimodal vs Unimodal
Dependent Variable	Primary performance metric used to evaluate model effectiveness	Accuracy
Moderator	Factors that influence performance variation across studies	Model, Sensor, Domain
Task	Types of tasks addressed in AIoT-based perception systems	Classification, Detection
Data Source	Databases used for collecting and selecting studies	Scopus, Wiley, ScienceDirect

This framework guides the organization of variables in the subsequent analysis, where the independent variable distinguishes between multimodal approaches and unimodal baselines, and accuracy is used as the primary outcome because it is the most consistently reported metric across the selected studies. However, not all studies provide comparable metrics or explicit baselines,

which affects how results can be interpreted and compared, and in turn motivates the inclusion of moderator variables such as model architecture, sensor modality, and application domain.

Multimodal Learning

Multimodal learning refers to the integration of information from multiple modalities, such as visual, physiological, or environmental data, into a shared representation. Different strategies are used to combine modalities, including data-level, feature-level, and decision-level fusion, as well as mechanisms that allow interaction across modalities (P. Chen et al., 2025; Ge et al., 2025; Xu et al., 2023; Y. Zhang et al., 2021). Across studies, improvements in accuracy are often observed when the modalities provide complementary information. However, this improvement is not consistent in all cases. The outcome depends on how well the modalities are aligned and whether the model can handle differences in data quality, synchronization, and distribution. When these aspects are not properly addressed, multimodal integration does not always lead to clear gains (Majumder & Kehtarnavaz, 2021; Y. Zhang et al., 2021).

Sensor Fusion

Sensor fusion focuses on combining data from multiple sensors to improve perception compared to using a single modality. Common strategies include early fusion, late fusion, and hybrid approaches, each with different implications for how information is integrated within the model (Xu et al., 2023; Y. Zhang et al., 2021). The effectiveness of sensor fusion depends on the relationship between the modalities being combined. Complementary modalities tend to produce stronger improvements, while noisy or misaligned modalities can reduce performance if the model cannot account for their reliability. This explains why reported gains vary across studies and are not always directly comparable (Ge et al., 2025; Majumder & Kehtarnavaz, 2021).

Computer Vision in AIoT

Computer vision is widely used in AIoT systems as a primary source of contextual information, particularly for classification and detection tasks (Lin, 2023; Y. Zhang et al., 2021). These studies primarily focus on classification and detection tasks as common benchmarks for evaluating perception performance in AIoT systems. However, relying on vision alone introduces limitations, especially in environments with noise, occlusion, or changing conditions. These limitations are often addressed by combining vision with other modalities. In the reviewed studies, vision is frequently integrated with additional data sources such as clinical data, physiological signals, or IoT sensors. This combination is used to improve classification performance across different domains. At the same time, differences in datasets, evaluation metrics, and reporting practices make it difficult to compare results across studies, particularly when baseline values are not consistently provided.

Evaluation Challenges in AI-Based Systems

One recurring issue across the literature is the lack of standardized evaluation practices. Differences in baselines, datasets, and reporting methods make it difficult to determine whether observed improvements are generalizable or limited to specific experimental conditions (Ayetiran & Özgöbek, 2024; L. Wu et al., 2023; Z. Zhang et al., 2020). In addition, variations in model design, sensor combinations, and application domains contribute to heterogeneous outcomes. These variations suggest that performance is not determined by a single factor, but by the interaction between multiple components. Understanding these relationships is necessary for interpreting differences across studies and for designing systems that perform reliably in real-world AIoT environments.

e-ISSN 2828-0490
JURNAL MEDIA COMPUTER SCIENCE

Research Methodology

This study was conducted using a combination of systematic review and meta-analysis to compare multimodal approaches with unimodal baselines in AIoT-based perception tasks. The review process broadly follows established reporting practices for systematic reviews, although it was adapted to accommodate common limitations in AI literature, particularly the lack of consistent baselines and comparable evaluation metrics across studies (Casarramona et al., 2022; Page et al., 2021).

The study collection was carried out across three major databases: Scopus, Wiley Online Library, and ScienceDirect. The search query combined terms related to multimodal learning, sensor fusion, computer vision, and accuracy in IoT contexts. The initial search yielded 362 articles. An early filtering stage was then applied based on publication year (2020–2025), document type (journal articles), and accessibility (open access), reducing the dataset to 76 studies.

Subsequent screening was more selective. Title and abstract screening reduced the number of studies from 76 to 16, as many articles addressed IoT or vision independently without incorporating multimodal integration. A refined screening stage further narrowed the selection to 14 studies. During full-text assessment, one study was excluded due to inaccessibility, resulting in 13 studies for final inclusion. This process reflects the challenge of retrieving fully relevant studies using broad keyword-based queries. The overall selection workflow is summarized in Figure 1.

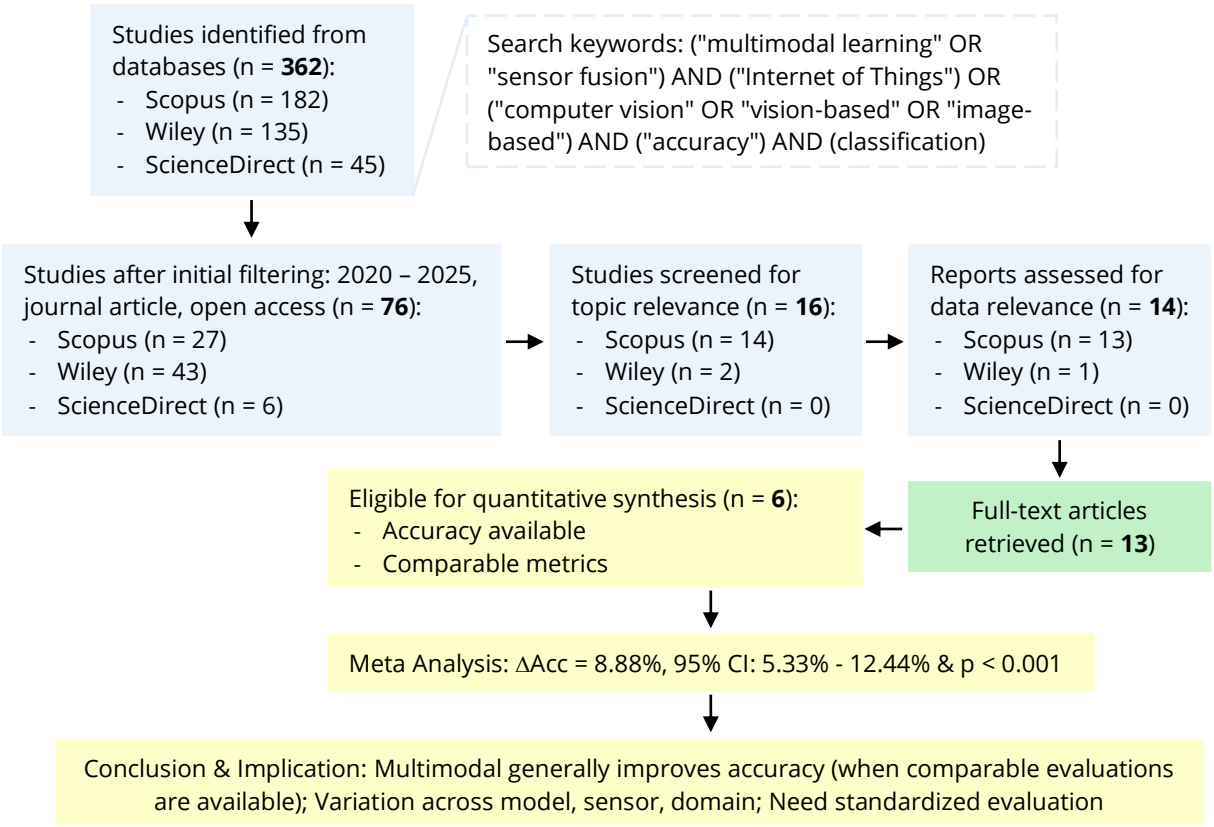

Figure 1. Study Selection Process and Meta-Analysis Workflow of Multimodal versus Unimodal Approaches in AIoT Systems

At this point, a more significant issue became clear. Although all 13 studies were relevant in a general sense, not all of them could be used for quantitative synthesis. Several studies reported only multimodal results without providing a clear unimodal baseline. Others used evaluation metrics that

were not directly comparable. Because of this, the dataset had to be split. All 13 studies were retained for qualitative analysis, while only 6 studies could be used in the meta-analysis.

Data extraction focused on a small set of variables that were consistently available across studies. These included baseline accuracy (AccC), multimodal accuracy (AccE), model type, sensor modality, and application domain. In practice, even these variables were not always reported in a clean or standardized way. Some papers required interpretation to identify what could reasonably be treated as a baseline. This introduces a level of approximation, but excluding such studies entirely would further reduce the already limited dataset. The extracted data are summarized in Table 2 and Table 3. Table 2 includes studies with comparable baseline and multimodal performance used for meta-analysis, while Table 3 presents studies excluded from quantitative synthesis along with the reasons.

Table 2. Extracted Data for Meta-Analysis Including Baseline and Multimodal Performance

Study	Performance (AccC / AccE / ΔAcc)	Model (Baseline / Multimodal)	Context (Sensor / Task / Dataset / Notes)
S01 (Njoroge et al., 2025)	88.4%; 95.8% → 99.2% (+10.8%; +3.4%)	DenseNet50; ShuffleNet → Hybrid DL fusion	Vision + IoT / Classification / Kaggle + FieldPlant / Valid
S02 (Jahani et al., 2024)	65.1%; 72% → 76.9% (+11.8%; +4.9%)	3D-DenseNet → 3D-DenseNet fusion	sMRI + fMRI / Classification / ABIDE / Valid
S03 (Ishida et al., 2024)	69.9%; 77.5% → 84.1% (+14.2%; +6.6%)	XGBoost; Autoencoder → Hybrid	Image + clinical / Classification / Internal / Valid
S05 (John et al., 2021)	94.12%; 85.26% → 99.72% (+5.60%; +14.46%)	1D-CNN → CNN fusion	ECG + SpO2 / Classification / UCD / Valid
S10 (Kansana et al., 2025)	97.40%; 96.70%; 95.60% → 99.40%; 100% (+2.00%; +2.70%; +3.80%)	Transformer; XGBoost; RF → Transformer fusion	Vision + tactile / Classification / Touch and Go / Valid
S12 (Cai et al., 2022)	86.66% → 95.55% (+8.89%)	SOLOv1 → Fusion CNN	RGB + hyperspectral / Classification / Custom / Valid

Table 3. Summary of Excluded Studies From Quantitative Synthesis and Reasons for Exclusion (S01–S13 Denote the Indexed Studies Selected After Full-Text Screening)

Study	Performance (AccC / AccE / ΔAcc)	Model (Baseline / Multimodal)	Context (Sensor / Task / Dataset / Notes)
S04 (Partel et al., 2021)	84% → —	CNN → Sensor fusion system	LiDAR + camera / Detection + classification / Internal / Multimodal performance (AccE) not reported, preventing effect size estimation
S06 (Rashid et al., 2023)	— → 94.12%; 93.68%; 86.34%; 86.19%	— → Ensemble ML	Physiological sensors / Classification / WESAD / No explicit baseline under comparable experimental settings
S07 (Mekruksavanich & Jitpattanakul, 2025)	— → 97.32%; 97.14%; 98.68%	— → CNN-ResBiGRU	IMU sensors / Classification / VTT-ConIoT / Lack of paired baseline–multimodal comparison within a unified

Study	Performance (AccC / AccE / ΔAcc)	Model (Baseline / Multimodal)	Context (Sensor / Task / Dataset / Notes)
			evaluation setup
S08 (Sadhwani et al., 2025)	~60–80% → 98.04%; 94.70%; 86.63%	CNN, ViT, YOLO → VLM	Vision + Text / Classification / Multi dataset / Heterogeneous datasets and models without consistent paired baseline comparison
S09 (Mwale et al., 2025)	0.861; 0.852 → 0.925; 0.886	LSTM; CNN → CNN-LSTM	Multisensor + vision / Classification / Mining dataset / Multi-task framework with heterogeneous evaluation metrics, lacking a unified comparable outcome
S11 (Y. Chen et al., 2021)	89%; ~98% → 93–99.38%	CNN → CNN fusion	Vision + proprioception / Classification / Internal / Performance reported as ranges across multiple conditions without consistent paired comparison
S13 (Gutiérrez-Ramírez et al., 2025)	— → 98.62%	— → ANN fusion	Vision + sensor / Classification / Internal / Baseline performance not explicitly reported under the same experimental setting

This separation reflects inconsistencies in reporting practices, where several studies do not provide sufficient information for direct comparison. For example, study S08 reports baseline performance as a range and involves multiple models, making it unclear which values should be paired for comparison. In another case, study S09 reports numerical results that appear comparable at first glance, but the evaluation metrics and pairing between baseline and multimodal results are not clearly defined. In several other studies, multimodal results were reported without a clearly defined baseline, making it difficult to determine the actual improvement. These issues were not rare, and they appeared repeatedly during the screening and extraction process, which ultimately reduced the number of studies that could be included in the meta-analysis.

The meta-analysis was conducted using OpenMEE, with the effect size defined as the difference in accuracy between multimodal and unimodal approaches (ΔAcc = AccE − AccC). A random-effects model was used, given the clear variability across datasets, models, and experimental setups, which is commonly recommended when between-study heterogeneity is expected (Casarramona et al., 2022; Olgun et al., 2025; Ren et al., 2024). One practical limitation was the absence of reported variance or standard deviation in most AI studies. To proceed, a constant variance value of 1 was assigned to all effect sizes. This simplification deviates from standard variance estimation approaches such as DerSimonian–Laird or REML, but was necessary due to missing statistical reporting in the selected studies (Ren et al., 2024). This is a simplification, and it should be interpreted accordingly, but it allows aggregation in situations where statistical reporting is incomplete. Without this adjustment, most of the collected studies could not be included in the analysis.

The statistical analysis focused on estimating the overall effect size and its 95% confidence interval, mainly to see whether the observed improvement was consistent across studies rather than driven by a few extreme results. Heterogeneity was examined using the I^2 statistic, alongside Cochran's Q, to capture variability across studies, which are standard measures for assessing between-study differences in meta-analysis (Casarramona et al., 2022; Guan et al., 2025; Smajic et al., 2025). Given the diversity in study design, a high level of heterogeneity was anticipated and later confirmed in the results. In such cases, additional analyses such as prediction intervals or sensitivity checks are often recommended to better reflect uncertainty beyond the average effect (Ren et al.,

2024; Smajic et al., 2025). This step was necessary to understand whether differences between studies were systematic rather than random. It also provides context for interpreting subgroup analysis in the next stage.

To check whether a small number of studies were disproportionately influencing the results, additional analyses were carried out. These included cumulative analysis, leave-one-out analysis, and fail-safe N, which are commonly used to assess robustness and the influence of individual studies in meta-analytic settings (Casarramona et al., 2022; Olgun et al., 2025; Smajic et al., 2025). Each of these approaches examines stability from a different angle, either by adding studies sequentially or by removing them one at a time. Together, these steps provide a more complete picture of how stable the findings are, even when the dataset is relatively small. This is particularly relevant given the limited number of studies included in the meta-analysis. Overall, the workflow moved from broad collection to a much narrower set of comparable studies. The process is not perfectly clean, mainly due to inconsistencies in how AI studies report results. Instead of forcing uniformity, the methodology keeps these limitations visible and incorporates them into the analysis. This reflects the actual state of the literature, where comparison is possible, but only within a constrained subset of studies.

RESULTS AND DISCUSSION

Results

Data Overview and Effect Size Results

A total of 13 studies (S01 - S13) were included in the analysis, all of which were retained for qualitative examination, although only 6 provided data that could be reasonably aligned for quantitative synthesis. This imbalance became more visible during data extraction, where baseline performance was not always explicitly reported and evaluation metrics were sometimes defined in ways that made direct comparison difficult. In several cases, results initially appeared comparable, but closer inspection showed that the pairing between unimodal and multimodal performance was not clearly defined, making them unsuitable for inclusion in the meta-analysis. As a result, part of the literature had to be set aside from the quantitative synthesis, even though these studies remained relevant for understanding broader patterns.

This inconsistency reflects a broader issue across the analyzed studies, where reporting practices are often sufficient for demonstrating individual model performance but not for enabling cross-study comparison. Several studies (e.g., S06, S07, S08, S13) reported multimodal performance without clearly defined unimodal baselines, while others (e.g., S09, S11) used metrics or reporting formats that limited comparability. Consequently, the quantitative findings rely on a smaller but more consistent subset of studies, while the remaining works contribute to qualitative interpretation.

The meta-analysis was conducted using a random-effects model (DerSimonian-Laird) to account for variability across studies. The pooled effect size, defined as the difference in accuracy between multimodal and unimodal approaches, was estimated at 8.882 with a 95% confidence interval ranging from 5.328 to 12.435 and a statistically significant p-value below 0.001. These results indicate that, where comparable evaluations are available, multimodal approaches tend to outperform unimodal baselines, with an average improvement of approximately 8.88%, consistent with prior findings that gains are task-, data-, and method-dependent (Bayoudh et al., 2021; Jiao et al., 2024). Importantly, this improvement should not be interpreted as uniform across all conditions. The spread of the confidence interval suggests that performance gains vary depending on study-specific factors. This pattern is consistent across the included studies (S01- S13), where, when comparable evaluations are available, multimodal approaches demonstrate improvement, but with varying magnitudes depending on domain, sensor configuration, and model design.

e-ISSN 2828-0490
JURNAL MEDIA COMPUTER SCIENCE

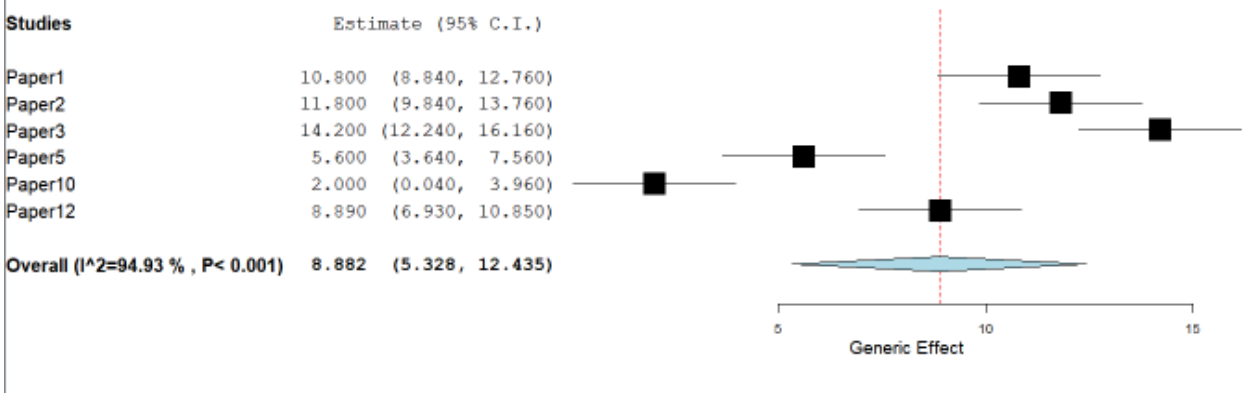

Figure 2. Forest Plot of Overall Meta-Analysis Results

As illustrated in Figure 2, all individual study estimates are located on the positive side of the reference line, indicating that multimodal approaches consistently outperform unimodal baselines. However, the dispersion of the estimates reflects substantial variability across studies. This variation suggests that while multimodal learning provides a consistent advantage, the magnitude of improvement is conditional rather than universal, aligning with evidence that modality quality, alignment, and domain characteristics mediate gains (Bayoudh et al., 2021; Jiao et al., 2024).

Variability, Subgroup Patterns, and Robustness

The variability across studies is substantial, as indicated by the high heterogeneity (I^2 = 94.929%). This level of heterogeneity suggests that most of the observed variation is not due to random error, but rather to differences in study characteristics, including domain, sensor modalities, and model architectures. This aligns with the diversity observed across the analyzed studies (S01 - S13), where multimodal systems are implemented in different contexts such as healthcare, agriculture, industrial monitoring, and smart environments, and where domain shift and dataset characteristics are known drivers of performance variability (Bayoudh et al., 2021; Jiao et al., 2024). To further examine this variability, subgroup analyses were conducted based on application domain, sensor configuration, and model architecture.

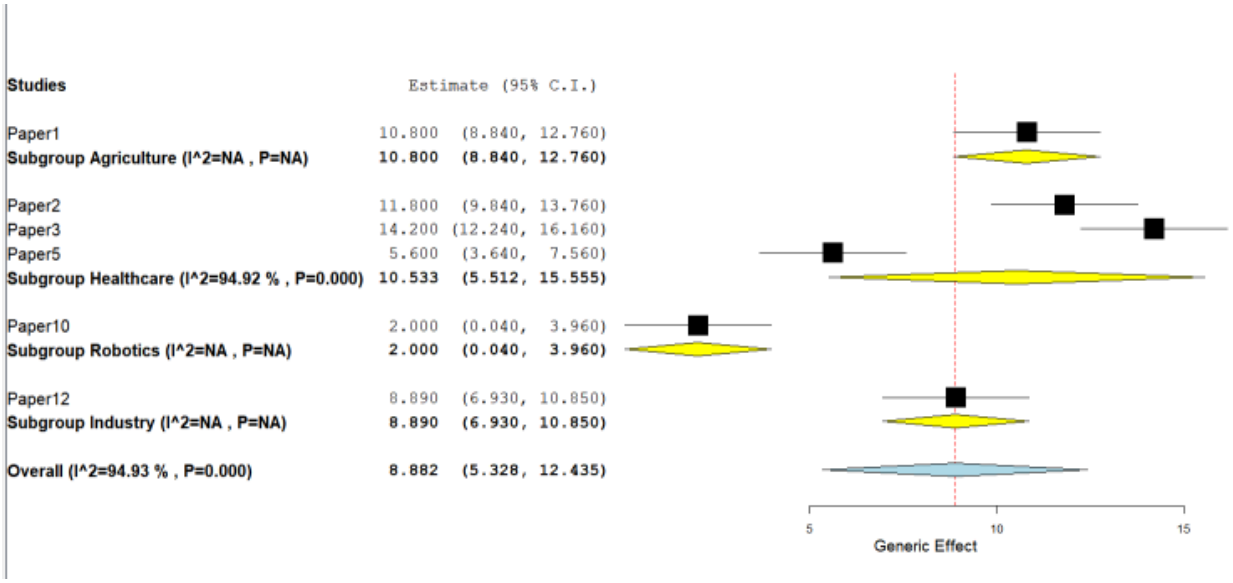

Figure 3. Subgroup Analysis Based on Application Domain

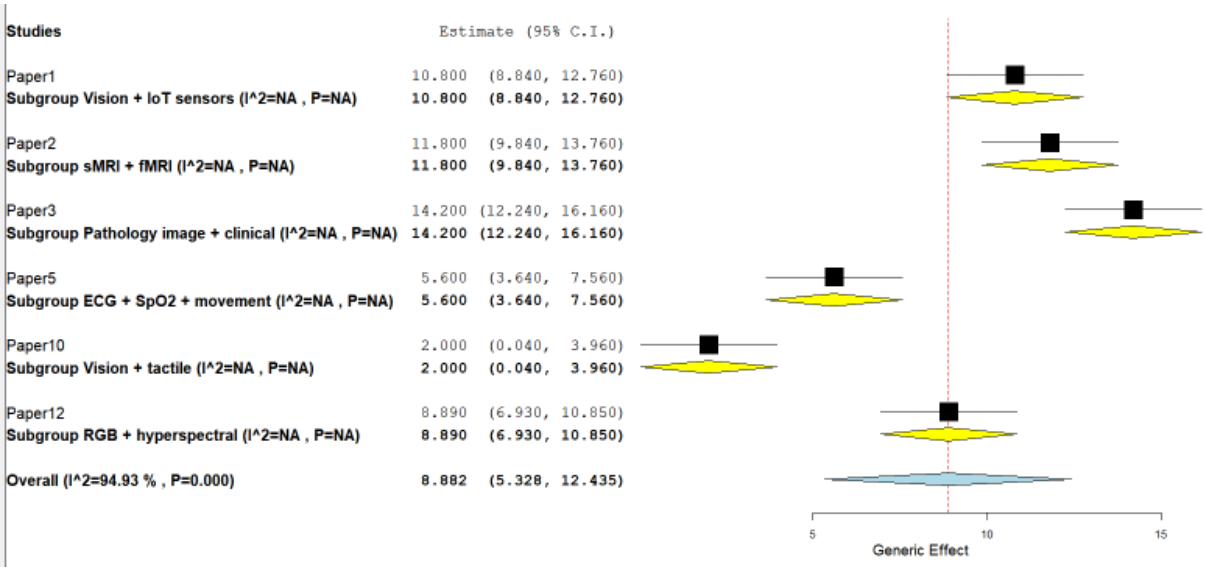

Figure 4. Subgroup Analysis Based on Sensor Modality Combinations

The domain-based analysis shows that clinically structured studies tend to exhibit higher performance gains compared to other domains. Studies such as S02 and S03 demonstrate that combining complementary modalities (e.g., structural and functional data) leads to more stable improvements, while S05 reflects physiological sensing in wearable settings rather than clinical structured data. In contrast, domains such as robotics or real-time monitoring (e.g., S04, S07) show more moderate gains, likely due to higher noise levels and real-time constraints. The analysis of sensor configurations indicates that heterogeneous modality combinations tend to produce stronger improvements than homogeneous ones, consistent with the role of complementarity and information diversity in enhancing discriminative power (Bayoudh et al., 2021; Jiao et al., 2024). Studies such as S01, S03, and S10 demonstrate that combining modalities with distinct characteristics leads to more substantial performance gains. Conversely, adding similar or redundant modalities does not necessarily improve performance and may even introduce noise, as observed in S05 and S06.

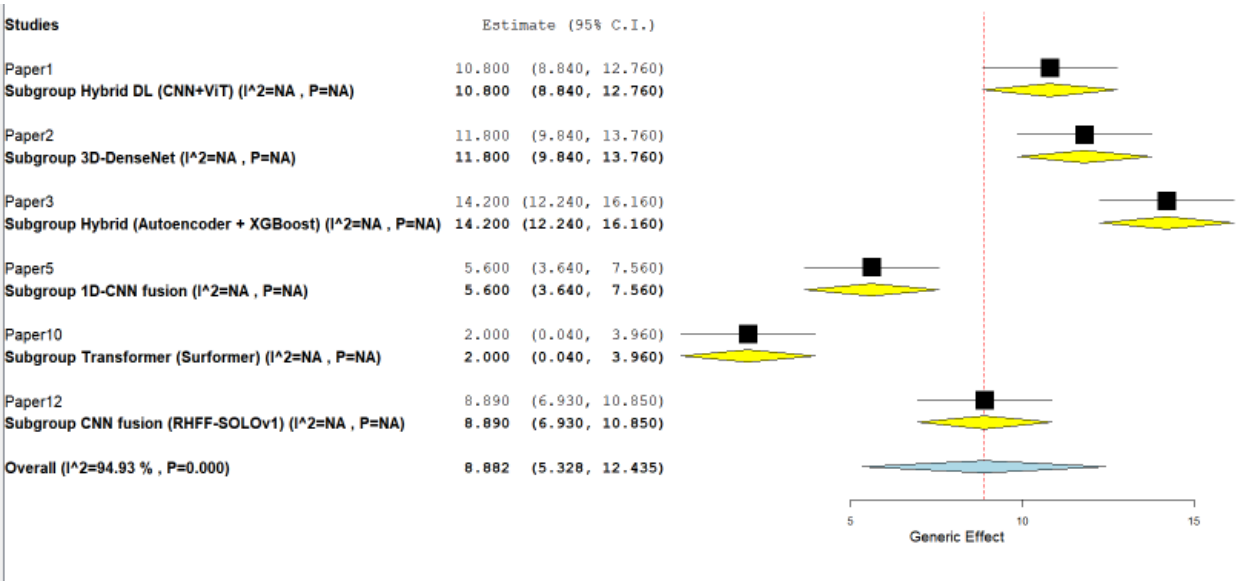

Figure 5. Subgroup Analysis Based on Model Architecture

e-ISSN 2828-0490
JURNAL MEDIA COMPUTER SCIENCE

Across model architectures, hybrid models consistently achieve higher effect sizes compared to single-model approaches, in line with evidence that fusion strategy (e.g., attention-based or hybrid integration) critically shapes multimodal effectiveness (Bayoudh et al., 2021; Jiao et al., 2024). Studies such as S03 and S10 show that combining different learning mechanisms enables better capture of cross-modal relationships. This suggests that model design plays a critical role in determining the effectiveness of multimodal systems. To assess the stability of these findings, cumulative and leave-one-out analyses were conducted.

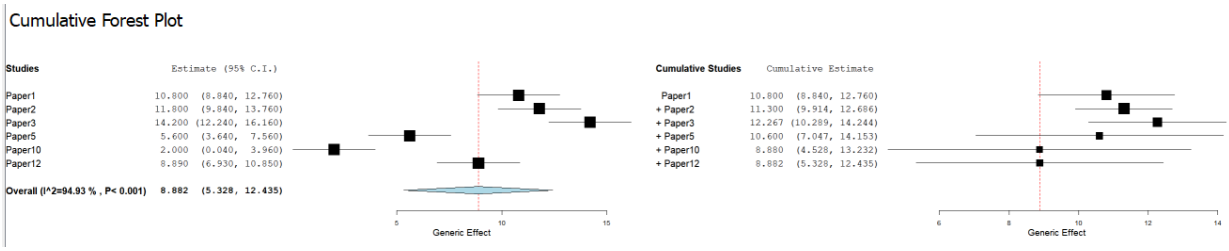

Figure 6. Cumulative Forest Plot Analysis

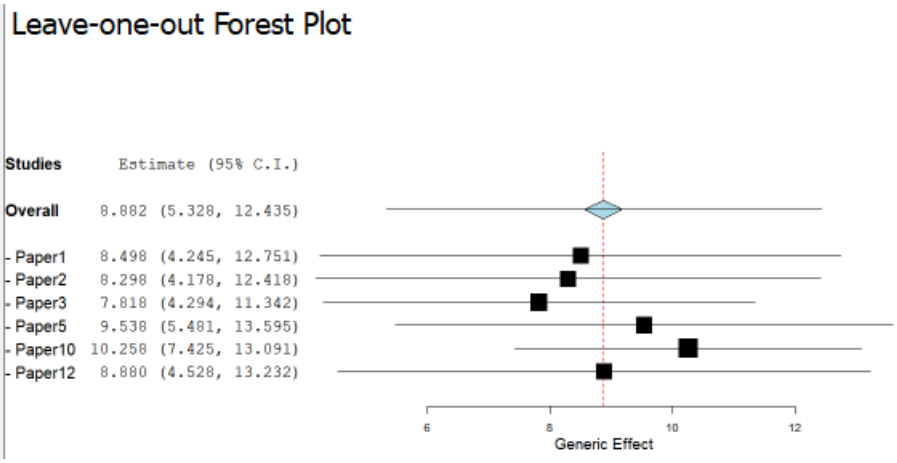

Figure 7. Leave-One-Out Sensitivity Analysis

The cumulative analysis indicates that the pooled effect size converges as more studies are added, suggesting increasing stability of the estimate. Similarly, the leave-one-out analysis shows that removing individual studies does not substantially alter the overall effect size, which remains within a narrow range. These results indicate that the observed improvement is robust and not driven by a small subset of studies. Table 4 summarizes the key quantitative findings, including effect size, heterogeneity, and robustness indicators.

Table 4. Summary of Meta-Analysis Results

Component	Metric	Result	Interpretation
Overall Effect	ΔAcc	8.882	Moderate improvement in accuracy
Confidence Interval	95% CI	5.328 – 12.435	Statistically significant range
Significance	p-value	< 0.001	Strong evidence against null hypothesis
Heterogeneity	I^2	94.929%	High variability across studies
Heterogeneity	Q	98.608	Significant between-study variation
Variance	τ^2	18.722	Substantial dispersion in effect size

Component	Metric	Result	Interpretation
Robustness	Fail-safe N	1044	Complementary robustness indicator
Robustness	Leave-one-out	7.818 – 10.258	Stable across study removal
Robustness	Cumulative	Convergent	Effect stabilizes with added studies

Thematic Synthesis Framework for Multimodal Learning

To better connect the quantitative results with their underlying interpretation, a thematic synthesis was developed based on the qualitative analysis of all included studies (S01-S13). This synthesis organizes the key determinants of multimodal performance into four overarching dimensions. As summarized in Table 5, these themes provide a structured explanation of the observed heterogeneity in the meta-analysis, capturing conditional performance gains, modality complementarity, fusion design, and practical trade-offs.

Table 5. Thematic Synthesis of Multimodal Learning in AIoT Systems

Theme	Description	Supporting Studies	Key Insight
Conditional Effectiveness	Multimodal improves performance but not uniformly	(Ishida et al., 2024; Mwale et al., 2025; Njoroge et al., 2025; Rashid et al., 2023)	Performance depends on context
Complementarity & Interaction	Modalities must provide complementary information	(Ishida et al., 2024; Jahani et al., 2024; Mekruksavanich & Jitpattanakul, 2025; Partel et al., 2021)	Not all modalities contribute equally
Fusion Strategy	Performance is driven by how modalities are fused	(Cai et al., 2022; John et al., 2021; Kansana et al., 2025; Sadhwani et al., 2025)	Fusion design determines outcome
Practical Constraints	Real-world deployment introduces trade-offs	(Y. Chen et al., 2021; Gutiérrez-Ramírez et al., 2025; Mwale et al., 2025; Partel et al., 2021)	Accuracy must be balanced with efficiency

Discussion

Multimodal Learning in AIoT: From Performance Gain to System Design

This study advances a conceptual perspective on multimodal learning in AIoT, where performance gains are not inherently guaranteed but are conditional on three interrelated factors: (i) modality complementarity, (ii) fusion strategy, and (iii) system-level constraints. These factors jointly determine the extent to which multimodal integration can effectively enhance system performance, reframing multimodal learning from a purely performance-driven paradigm into a system design problem, consistent with prior syntheses emphasizing complementarity, fusion design, and domain-dependent variability.

e-ISSN 2828-0490
JURNAL MEDIA COMPUTER SCIENCE

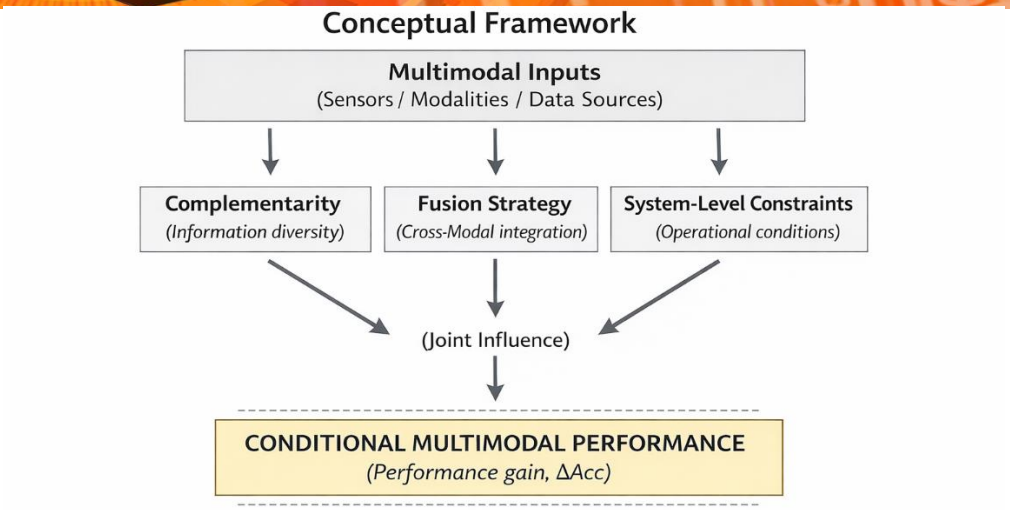

Figure 8. Conceptual Framework for Conditional Multimodal Performance

To further formalize the conceptual framework illustrated in Figure 8, Table 6 summarizes its key dimensions, their roles, and their implications for multimodal system design. This formalization provides a structured lens for interpreting how different design factors contribute to performance variability across studies. It also highlights that multimodal effectiveness should be evaluated not only in terms of accuracy gains, but in relation to system design choices and operational context.

Table 6. Conceptual Dimensions of Multimodal AIoT Effectiveness

Dimension & Key Question	Role in System	Impact on Performance	Implication for Design
Complementarity (What is combined?)	Determines information diversity	Enables meaningful performance gain	Select modalities with non-redundant and complementary information
Fusion Strategy (How is it combined?)	Governs interaction and representation learning	Controls effectiveness of integration	Choose fusion techniques aligned with data heterogeneity and task requirements
System Constraints (Where is it deployed?)	Defines operational limitations	Limits or amplifies achievable gains	Balance accuracy with efficiency, latency, and resource constraints
Outcome (What is achieved?)	Resulting system performance	Conditional (not universally guaranteed)	Evaluate performance within system context rather than absolute metrics

Conditional Effectiveness of Multimodal Learning

The results indicate that multimodal approaches generally improve performance compared to unimodal baselines when comparable evaluations are available. However, this improvement is not uniform across all conditions. While studies such as S01, S02, S03, S05, and S10 show clear and significant gains, others indicate that the magnitude of improvement depends on specific system configurations. This is reflected in the high heterogeneity observed in the meta-analysis, consistent with evidence that multimodal gains vary under data heterogeneity, noise, and modality availability (Hui et al., 2025; Shin et al., 2025).

These findings suggest that multimodal effectiveness should be understood as conditional rather than universal. Performance gains depend on how modalities are combined, the nature of the task, and the characteristics of the data. For example, in structured domains such as healthcare

(S02, S03), multimodal integration tends to produce more consistent improvements, whereas in dynamic or real-time environments (S04, S07), gains may be more variable.

Role of Complementarity and Modality Interaction

A key factor underlying multimodal performance is the complementarity between modalities, where information diversity rather than redundancy drives gains and robustness (Bayoudh et al., 2021). Across the analyzed studies, improvements are generally observed when modalities provide distinct but related information. For instance, S02 combines structural and functional data, S03 integrates imaging with clinical features, and S10 fuses heterogeneous sensor signals.

However, not all modalities contribute equally. Several studies (e.g., S02, S03, S07, S05, S06) indicate that certain modalities dominate others in terms of contribution. In some cases, adding more modalities does not lead to proportional improvements and may even reduce performance due to noise or redundancy. This highlights the importance of selecting modalities based on relevance rather than quantity, aligning with prior work showing that complementarity, rather than redundancy, drives robust gains (Liu et al., 2021; Y. Wu et al., 2025).

Fusion Strategy as the Main Performance Driver

Beyond modality selection, the way modalities are fused plays a central role in determining system performance, with the effectiveness of early, intermediate, and late fusion depending on task and data characteristics (Jiao et al., 2024). The analyzed studies demonstrate a range of fusion strategies, including feature-level fusion (S03, S05), system-level sensor fusion (S04), hybrid architectures (S10), and ensemble-based multimodal integration (S06). These variations represent fundamentally different approaches to multimodal integration. Representation-based fusion enables alignment of heterogeneous data, while adaptive fusion dynamically selects modalities based on context. Hybrid models combine different learning mechanisms to capture complex relationships. This diversity in fusion strategies explains a substantial portion of the heterogeneity observed in the meta-analysis, indicating that system design plays a more critical role than modality count alone, particularly the choice of fusion strategy (e.g., early vs. late vs. model-level fusion) and cross-modal interaction mechanisms (Che et al., 2025; Liu et al., 2021).

Practical Constraints and Real-World Deployment

In addition to performance considerations, several studies highlight practical constraints associated with multimodal systems. Real-world deployment introduces challenges related to computational cost, latency, sensor reliability, and environmental variability. For instance, S04 demonstrates real-time constraints in sensor fusion systems, while S10 emphasizes industrial deployment requirements. Similarly, S05 and S06 highlight trade-offs between sensor usage, energy consumption, and performance.

These constraints indicate that multimodal systems must balance accuracy with efficiency and robustness, as also reported in domains such as autonomous systems and remote sensing where latency, alignment, and energy constraints shape design choices (Hui et al., 2025; Li et al., 2023; Ngo et al., 2024; Wang, 2025). High-performing models may achieve superior accuracy but require greater computational resources, limiting applicability in edge environments. Therefore, effectiveness should be evaluated within the broader context of system requirements.

Limitations and Future Directions

Several limitations should be considered when interpreting these findings. A major constraint is the limited number of studies that could be included in the quantitative synthesis due to inconsistent reporting practices. Many studies lacked clear unimodal baselines or used incompatible evaluation metrics. In addition, the absence of variance measures required simplified assumptions, which may affect the precision of the estimated effect size. The high heterogeneity further indicates substantial variation in study design and context.

Future research should focus on improving reporting standards, particularly by providing explicit baselines and consistent evaluation metrics, addressing known limitations such as heterogeneous metrics, missing baselines, and limited reproducibility in AI studies (Bayoudh et al., 2021; Jiao et al., 2024). There is also a need for controlled studies that systematically compare unimodal and multimodal approaches. Furthermore, exploring adaptive and context-aware fusion strategies may provide more robust solutions for real-world AIoT systems. Overall, this study provides evidence that multimodal learning improves performance, but its effectiveness is shaped by system design, modality interaction, and practical constraints.

CONCLUSION AND SUGGESTIONS

Conclusion

This study provides a meta-analytic evaluation of multimodal learning in Artificial Intelligence of Things (AIoT) systems, focusing on the integration of sensor fusion and computer vision for classification tasks. Based on the analysis of 13 selected studies, with 6 studies included in the quantitative synthesis, the results indicate that multimodal approaches generally improve accuracy compared to unimodal baselines when comparable evaluations are available.

The meta-analysis shows an average performance improvement of approximately 8.88%, with statistically significant results. However, the high heterogeneity across studies demonstrates that this improvement is not uniform and is influenced by multiple factors, including application domain, sensor modality, and model architecture. Subgroup analysis further reveals that structured domains such as healthcare and hybrid model architectures tend to yield more consistent and higher performance gains.

Beyond quantitative findings, this study highlights that multimodal effectiveness should be understood as conditional rather than universal. Performance gains depend on the complementarity of modalities, the design of fusion strategies, and system-level constraints such as computational resources and deployment environments. In this context, multimodal learning is better interpreted as a system design problem rather than merely a performance enhancement technique.

A key contribution of this study is the identification of a recurring limitation in the literature, namely the lack of standardized baseline reporting and comparable evaluation metrics. This limitation significantly reduces the number of studies that can be included in meta-analysis and affects the generalizability of conclusions. Overall, this study clarifies that while multimodal approaches offer measurable benefits, their effectiveness is shaped by context, design choices, and reporting quality.

Suggestions

Based on the findings and limitations identified in this study, future research should prioritize the standardization of evaluation practices, particularly by providing explicit unimodal baselines and consistent performance metrics to enable reliable cross-study comparison and meta-analysis. In addition, multimodal system development should be approached from a design-oriented perspective, where the selection of modalities is guided by complementarity rather than quantity, ensuring that each modality contributes meaningful and non-redundant information. There is also a need to explore adaptive and context-aware fusion mechanisms that can dynamically respond to variations in data quality, sensor reliability, and environmental conditions, thereby improving robustness in real-world deployments.

At the same time, practical system constraints such as latency, computational cost, and energy efficiency should be more explicitly integrated into model design, particularly for edge-based AIoT applications where resource limitations are critical. Future studies are also encouraged to conduct

more controlled experimental comparisons between unimodal and multimodal approaches under consistent settings. Such efforts are essential to strengthen empirical evidence, improve reproducibility, and enhance the generalizability of findings in multimodal AIoT research.

REFERENCES

Ayetiran, E. F., & Özgöbek, Ö. (2024). A Review of Deep Learning Techniques for Multimodal Fake News and Harmful Languages Detection. Ieee Access, 12, 76133–76153. https://doi.org/10.1109/access.2024.3406258

Bayoudh, K., Knani, R., Hamdaoui, F., & Mtibaa, A. (2021). A Survey on Deep Multimodal Learning for Computer Vision: Advances, Trends, Applications, and Datasets. The Visual Computer, 38(8), 2939–2970. https://doi.org/10.1007/s00371-021-02166-7

Cai, Z., Yang, J., Fang, H., Ji, T., Hu, Y., & Wang, X. (2022). Research on Waste Plastics Classification Method Based on Multi-Scale Feature Fusion. Sensors, 22(20), 7974. https://doi.org/10.3390/s22207974

Casarramona, G. L., Lalmahomed, T. A., Lemmen, C. H. C., Eijkemans, M. J. C., Broekmans, F. J. M., Cantineau, A. E. P., & Drechsel, K. (2022). The Efficacy and Safety of Luteal Phase Support With Progesterone Following Ovarian Stimulation and Intrauterine Insemination: A Systematic Review and Meta-Analysis. Frontiers in Endocrinology, 13. https://doi.org/10.3389/fendo.2022.960393

Che, J., Sun, M., Wang, Y., & Xu, Z. (2025). Fusion-Driven Multimodal Learning for Biomedical Time Series in Surgical Care. Frontiers in Physiology, 16. https://doi.org/10.3389/fphys.2025.1605406

Chen, P., Zhao, X., Zeng, L., Liu, L., Liu, S., Li, S., Li, Z., Chen, H., Liu, G., Qiao, Z., Qu, Y., Xu, D., Li, L., & Li, L. (2025). A Review of Research on SLAM Technology Based on the Fusion of LiDAR and Vision. Sensors, 25(5), 1447. https://doi.org/10.3390/s25051447

Chen, Y., Rastogi, C., & Norris, W. R. (2021). A CNN Based Vision-Proprioception Fusion Method for Robust UGV Terrain Classification. IEEE Robotics and Automation Letters, 6(4), 7965–7972. https://doi.org/10.1109/LRA.2021.3101866

Ge, M., Ohtani, K., Niu, Y., Zhang, Y., & Takeda, K. (2025). VLA-MP: A Vision-Language-Action Framework for Multimodal Perception and Physics-Constrained Action Generation in Autonomous Driving. Sensors, 25(19), 6163. https://doi.org/10.3390/s25196163

Guan, X., Lan, M., Tang, L., Yang, H., Chen, Y., Ge, L., & Zhong, Y. (2025). Efficacy of Action Observation Therapy on Cognitive Function in Stroke: A Systematic Review and Meta-Analysis. Brain and Behavior, 15(4). https://doi.org/10.1002/brb3.70474

Gutiérrez-Ramírez, J. J., Macias-Jamaica, R. E., Zamudio-Rodríguez, V. M., Sotelo, H. A., Velázquez-Vázquez, D. A., de Anda-Suárez, J., & Gutiérrez-Hernández, D. A. (2025). A Modular Framework for RGB Image Processing and Real-Time Neural Inference: A Case Study in Microalgae Culture Monitoring. Eng, 6(9), 221. https://doi.org/10.3390/eng6090221

Hui, Q., Wang, M., Zhu, M., & Wang, H. (2025). A Review of Multi-Sensor Fusion in Autonomous Driving. Sensors, 25(19), 6033. https://doi.org/10.3390/s25196033

Ishida, S., Morita, K., Hatakeyama, K., Ren, N., Watanabe, S., Kobashi, S., Iihara, K., & Wakabayashi, T. (2024). Prediction of cardiovascular events after carotid endarterectomy using pathological images and clinical data. International Journal of Computer Assisted Radiology and Surgery, 20(4), 643–652. https://doi.org/10.1007/s11548-024-03286-w

Jahani, A., Jahani, I., Khadem, A., Braden, B. B., Delrobaei, M., & MacIntosh, B. J. (2024). Twinned neuroimaging analysis contributes to improving the classification of young people with autism spectrum disorder. Scientific Reports, 14(1), 20120. https://doi.org/10.1038/s41598-024-71174-z

Jiao, T., Guo, C., Feng, X., Chen, Y., & Song, J. (2024). A Comprehensive Survey on Deep Learning Multi-Modal Fusion: Methods, Technologies and Applications. Computers Materials & Continua, 80(1), 1–35. https://doi.org/10.32604/cmc.2024.053204

John, A., Nundy, K. K., Cardiff, B., & John, D. (2021). Multimodal Multiresolution Data Fusion Using Convolutional Neural Networks for IoT Wearable Sensing. IEEE Transactions on Biomedical Circuits and Systems, 15(6), 1161–1173. https://doi.org/10.1109/TBCAS.2021.3134043

Kansana, M., Hossain, E., Rahimi, S., & Amiri Golilarz, N. (2025). Surformer v1: Transformer-Based Surface Classification Using Tactile and Vision Features. Information, 16(10), 839. https://doi.org/10.3390/info16100839

Lei, L., Tan, Y., Zheng, K., Liu, S., Zhang, K., & Shen, X. (2020). Deep Reinforcement Learning for Autonomous Internet of Things: Model, Applications and Challenges. Ieee Communications Surveys & Tutorials, 22(3), 1722–1760. https://doi.org/10.1109/comst.2020.2988367

Li, W., Hacid, H., Almazrouei, E., & Debbah, M. (2023). A Comprehensive Review and a Taxonomy of Edge Machine Learning: Requirements, Paradigms, and Techniques. Ai, 4(3), 729–786. https://doi.org/10.3390/ai4030039

Lin, H.-Y. (2023). Embedded Artificial Intelligence: Intelligence on Devices. Computer, 56(9), 90–93. https://doi.org/10.1109/mc.2023.3280397

Liu, J., Chen, S., Wang, L., Liu, Z., Fu, Y., Guo, L., & Dang, J. (2021). Multimodal Emotion Recognition With Capsule Graph Convolutional Based Representation Fusion. 6339–6343. https://doi.org/10.1109/icassp39728.2021.9413608

Majumder, S., & Kehtarnavaz, N. (2021). Vision and Inertial Sensing Fusion for Human Action Recognition: A Review. Ieee Sensors Journal, 21(3), 2454–2467. https://doi.org/10.1109/jsen.2020.3022326

Mekruksavanich, S., & Jitpattanakul, A. (2025). Construction Worker Activity Recognition Using Deep Residual Convolutional Network Based on Fused IMU Sensor Data in Internet-of-Things Environment. IoT, 6(3), 36. https://doi.org/10.3390/iot6030036

Mwale, W., Liu, Z., & Chipusu, K. (2025). A Hybrid AI Framework for Integrated Predictive Maintenance and Mineral Quality Assessment in Mining. Applied Sciences, 15(22), 12222. https://doi.org/10.3390/app152212222

Ngo, N., Nguyen, K., Nazib, A., Fernando, T., Fookes, C., & Sridharan, S. (2024). Multimodal Colearning Meets Remote Sensing: Taxonomy, State of the Art, and Future Works. Ieee Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 17, 7386–7409. https://doi.org/10.1109/jstars.2024.3378348

Njoroge, T. K., Omol, E. J., & Nyangaresi, V. O. (2025). Deep Learning and IoT Fusion for Crop Health Monitoring: A High-Accuracy, Edge-Optimised Model for Smart Farming. IET Image Processing, 19(1). https://doi.org/10.1049/ipr2.70208

Olgun, N., Arayıcı, M. E., Kızmazoğlu, D., & Çeçen, E. (2025). Assessment of Chemo-Immunotherapy Regimens in Patients With Refractory or Relapsed Neuroblastoma: A Systematic Review With Meta-Analysis of Critical Oncological Outcomes. Journal of Clinical Medicine, 14(3), 934. https://doi.org/10.3390/jcm14030934

Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T., Mulrow, C. D., Shamseer, L., Tetzlaff, J., Akl, E. A., Brennan, S., Chou, R., Glanville, J., Grimshaw, J., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E., Mayo-Wilson, E., McDonald, S., … McKenzie, J. E. (2021). PRISMA 2020 Explanation and Elaboration: Updated Guidance and Exemplars for Reporting Systematic Reviews. BMJ, n160. https://doi.org/10.1136/bmj.n160

Partel, V., Costa, L., & Ampatzidis, Y. (2021). Smart tree crop sprayer utilizing sensor fusion and artificial intelligence. Computers and Electronics in Agriculture, 191, 106556. https://doi.org/10.1016/j.compag.2021.106556

Rashid, N., Mortlock, T., & Faruque, M. A. Al. (2023). Stress Detection Using Context-Aware Sensor Fusion From Wearable Devices. IEEE Internet of Things Journal, 10(16), 14114–14127. https://doi.org/10.1109/JIOT.2023.3265768

Ren, J., Ma, J., & Cappelleri, J. C. (2024). Appropriateness of Conducting and Reporting Random-effects Meta-analysis in Oncology. Research Synthesis Methods, 15(2), 326–331. https://doi.org/10.1002/jrsm.1702

Sadhwani, S., Shamsi, J. A., Khan, M. B., Bawany, N. Z., & Syed, H. J. (2025). Real-Time Detection of Mixed-Critical Events Using Vision-Language Models. IEEE Access, 13, 181363–181384. https://doi.org/10.1109/ACCESS.2025.3622638

Shin, J., Hassan, N., Miah, A. S. M., & Nishimura, S. (2025). A Comprehensive Methodological Survey of Human Activity Recognition Across Diverse Data Modalities. Sensors, 25(13), 4028. https://doi.org/10.3390/s25134028

Smajic, S., Konieczny, M. R., Kabir, K., Scrofani, R., Migliorini, F., & Dračić, A. (2025). Influence of Prone, Supine, and Lateral Positions During Spine Surgery on Vascular, Abdominal, and Postural Anatomy: A Comprehensive Review and Bayesian Meta-Analysis. European Journal of Medical Research, 30(1). https://doi.org/10.1186/s40001-025-03239-2

Utami, P., Mashoedah, Nurkhalis, H., Nafi', M. A., Savitri, W., Prastowo, W., Diah Wulan, S., & Saputra, F. D. (2024). An Integrated IoT–AI Architecture for Precision Beekeeping: Sensing, Data Communication, Colony-State Intelligence, and Decision-Oriented Actions. Jurnal Media Computer Science, 3(2), 179–198.

Utami, P., Zakarijah, M., Fikriawan, Z., Yanti, N., Aritonang, D., Gendhis Pertiwi, G., Rizqy Arba Pratama, A., & Azra, D. (2024). Iot-Enabled Dairy Systems: From Sensing And Data Integration To Operational Evaluation How to Cite. Jurnal Media Computer Science, 3(3), 199–216. https://doi.org/https://doi.org/10.37676/jmcs.v3i2.10590

Wang, L. (2025). Self-Supervised Learning and Transformer-Based Technologies in Breast Cancer Imaging. Frontiers in Radiology, 5. https://doi.org/10.3389/fradi.2025.1684436

Wu, D., Wu, S., Rawat, D. B., & Luo, C. (2024). Guest Editorial Special Issue on Cloud–Edge–Terminal Collaboration-Enabled AIoT: Services, Technologies, and Applications. Ieee Internet of Things Journal, 11(1), 1–4. https://doi.org/10.1109/jiot.2023.3331985

Wu, L., Liu, P., & Zhang, Y. (2023). See How You Read? Multi-Reading Habits Fusion Reasoning for Multi-Modal Fake News Detection. Proceedings of the Aaai Conference on Artificial Intelligence, 37(11), 13736–13744. https://doi.org/10.1609/aaai.v37i11.26609

Wu, Y., Mi, Q., & Gao, T. (2025). A Comprehensive Review of Multimodal Emotion Recognition: Techniques, Challenges, and Future Directions. Biomimetics, 10(7), 418. https://doi.org/10.3390/biomimetics10070418

Xu, P., Zhu, X., & Clifton, D. A. (2023). Multimodal Learning With Transformers: A Survey. Ieee Transactions on Pattern Analysis and Machine Intelligence, 45(10), 12113–12132. https://doi.org/10.1109/tpami.2023.3275156

Zhang, J., & Tao, D. (2021). Empowering Things With Intelligence: A Survey of the Progress, Challenges, and Opportunities in Artificial Intelligence of Things. Ieee Internet of Things Journal, 8(10), 7789–7817. https://doi.org/10.1109/jiot.2020.3039359

Zhang, Y., Sidibé, D., Morel, O., & Mériaudeau, F. (2021). Deep Multimodal Fusion for Semantic Image Segmentation: A Survey. Image and Vision Computing, 105, 104042. https://doi.org/10.1016/j.imavis.2020.104042

Zhang, Z., Sun, J., Li, Q., Liu, C., & Ding, G. (2020). A Novel MS-MeMBer Filter for Extended Targets Tracking. Ieee Access, 8, 37596–37607. https://doi.org/10.1109/access.2020.2975648