GURU KASHI UNIVERSITY
NAAC GRADE A++ ACCREDITED UNIVERSITY
INSTITUTION'S INNOVATION COUNCIL (Ministry of Education Initiative)

# GKU JOURNAL OF MULTIDISCIPLINARY RESEARCH

Volume - II, Issue - IV, October- 2026, ISSN:3049-298X

Peer - Reviewed Journal



Online ISSN : 3139 - 700X

Research ~ Innovation ~ Excellence

**Visit for more information**
https://journal.gku.ac.in

## Editorial Board

### Chief Editor

**Prof. (Dr.) Piush Verma,**
Pro Vice Chancellor R&D, Guru Kashi University, Talwandi Sabo, Bathinda, Punjab.

### Editor

**Prof. (Dr.) Vijay Laxmi**
Dean, Faculty of Computing, Guru Kashi University, Talwandi Sabo, Bathinda, Punjab.

### Co-Editor

**Dr. Rajinder Kumar**
Associate Professor, Faculty of Computing, Guru Kashi University, Talwandi Sabo, Bathinda, Punjab.

## Editorial Board Members

**Dr. Bhargav Koppolu**
Assistant Professor of Mechanical Engineering Program Director: Mechanical Engineering, Friends University, USA

**Ms. Mili Rahman**
Associate Professor, Department of English, Dhaka International University, Bangladesh

**Dr. Dheerendra Singh**
Associate Professor, Department of CSE, Chandigarh College of Engineering and Technology (CCET) Chandigarh

**Dr. Shruti Tiwari**
Principal, Gyan Vihar School of Education, Suresh Gyan Vihar University, Jaipur.

**Dr. Rajinder Kumar Uppal**
Professor & Principal, Guru Gobind Singh College of Management & Technology, Gidderbaha, Punjab

**Dr. Deepak Kumar**
Professor, Department of Computer Applications, Chandigarh University, Punjab

**Dr. Upinder Kaur**
Associate Professor and Head, Department of Computer Science and Engineering, Akal University, Talwandi Sabo, Bathinda, Punjab.

**Dr. Navneet Seth**
Dean and Professor, Faculty of Management and Commerce, Guru Kashi University, Talwandi Sabo Bathinda, Punjab

**Dr. Manpreet Kaur**
Assistant Professor, Faculty of Computing, Guru Kashi University, Talwandi Sabo, Punjab

## Special Guest Editors

**Dr. Amit Tuteja**
(Faculty of Sciences, Humanities & Languages)

**Dr. Rajinder Tiwari**
(Faculty of Engineering & Technology)

**Dr. Lokender Kashyap**
(Faculty of Agriculture)

**Dr. Satish Kumar**
(Faculty of Education)

**Dr. Rajesh**
(University School of Law)

**Dr. Devinder Kumar**
(Faculty of Pharmaceutical Sciences)

**Dr. Ravi Kumar**
(Faculty of Physical Education)

**Dr. Kanwaljit Kaur**
(Faculty of Visual & Performing Arts)

**Dr. Stuti Mittal**
(Faculty of Health & Allied Sciences)

**Dr. Sunny Arora**
(Faculty of Engineering & Technology)

**Prof. Rajender Kumar**
(Faculty of Health & Allied Sciences)

**Dr. Kavita Rani**
(Faculty of Education)

**Dr. Shalu Gupta**
(Faculty of Computing Science)

## AIM

The GKU Journal aims to serve as a distinguished platform for the publication and dissemination of high-quality, peer-reviewed research that contributes to the advancement of knowledge across multidisciplinary fields. The journal is dedicated to encouraging innovative research, critical inquiry, and scholarly excellence by providing researchers, academicians, professionals, and students with an avenue to share original ideas, research findings, and best practices.

The journal strives to foster intellectual collaboration, support evidence-based research, and promote ethical publishing practices while addressing emerging academic, technological, and societal issues. Through its commitment to quality and inclusivity, the GKU Journal seeks to make a meaningful contribution to the global research and academic community.

## SCOPE

The GKU Journal is a multidisciplinary, peer-reviewed academic publication dedicated to advancing high-quality research across diverse disciplines, including Science, Technology, Engineering, Management, Humanities, Social Sciences, Education, and allied fields. The journal welcomes original research articles, review papers, case studies, technical reports, short communications, and other scholarly contributions that demonstrate academic excellence and practical relevance.

The journal encourages research that addresses emerging global and societal challenges, promotes innovation, fosters interdisciplinary collaboration, and contributes to the generation of new knowledge. Emphasis is placed on studies that uphold the highest standards of academic integrity, ethical research practices, and meaningful societal impact.

As an open and inclusive platform for knowledge dissemination, the GKU Journal seeks to promote scholarly exchange among researchers, academicians, industry professionals, policymakers, and practitioners. It aims to strengthen collaboration between academia and industry while contributing to the advancement of national and international research communities.

# Index


| Sr. No. | Paper Title/Authors | Page No. |
|---|---|---|
| 1. | **Assessment of Holistic Intelligence among Boys Trainee Teachers at State Universities of Haryana**<br>o **Preeti Kukreja,** Research Scholar, Department of Law | 05-08 |
| 2. | **Cloud-based enterprise information systems are secure using zero trust architectures.**<br>o **Priyanka Sharma,** Assistant Professor, Department of Management | 09 – 13 |
| 3. | **Effect of Mindfulness Modules On Teaching Effectiveness of Prospective Teachers.**<br>o **Neema Nair,** Research Scholar, Department of Education<br>o **Dr. Harish Mittu,** Professor, Department of Education | 18 – 20 |
| 4. | **The Efficacy of Deep Belly Breathing (Adhama Pranayama) on Chronic Low Back Pain (CLBP): An Analytical Review**<br>o **Sandeep Kaur**, Research Scholar, Department of Physical Education<br>o **Sakshi Nagru** , Assistant Professor, Department of Physical Education. | 21 – 29 |
| 5. | **Perceived Benefits of Participation in Yogic Exercises between Rural and Urban University Students: A Comparative Study**<br>o **Rajveer Kaur,** Research Scholar, Department of Physical Education.<br>o **Sakshi Nagru** , Assistant Professor, Department of Physical Education. | 30 – 37 |
| 6. | **MOSAIC-Engage: Uncertainty-Aware Multimodal Deep Learning for Robust Student Engagement Estimation in Online Education**<br>o **Charanjeet Kaur,** Research Scholar, Department of Computer Application<br>o **Dr. Rajinder Kumar,** Associate Professor, Department of Computer Application . | 38 – 47 |
| 7. | **Artificial and Emotional Intelligence in Decision-Making: Evidence from the Indian Context Studies**<br>o **Sohail Verma,** Research Scholar, Department of Management<br>o **Dr. Pretty Bhala**, Assistant Professor Department of Management | 48 – 51 |
| 8. | **Machine Learning-Based Prediction of Hotel Guest Ratings: Price Associations and Geographical Transferability in Cairo, Egypt.**<br>o **Navjot Kaur,** Research Scholar, Department of Computer Application<br>o **Dr. Rajinder Kumar,** Associate Professor, Department of Computer Application. | 52 – 56 |
| 9. | **A Reproducibility-Focused Comparison of CNNs and Vision Transformers for Alzheimer's Disease Classification from Brain MRI**<br>o **Sahil Sharma,** Research Scholar, Department of Computer Application<br>o **Dr. Rajinder Kumar,** Associate Professor, Department of Computer Application. | 57 – 62 |
| 10. | **Biomechanical Analysis and Artificial Intelligence-Based Performance Assessment of Karate Punching Techniques: A Comprehensive Review**<br>o **Anija Mittal**, Research Scholar, Department of Computer Application<br>o **Dr. Shalu Gupta ,** Associate Professor, Department of Computer Application, | 63 – 68 |

# Assessment of Holistic Intelligence Among Boys Trainee Teachers at State Universities of Haryana.

Ms. Preeti Kukreja
Research Scholar, Department of Law, Lovely Professional University, Phagwara.
Email id: preeti.sethi80@gmail.com

Accepted:20-09-2026 Published: 12-10-2026 Page No. 4-8 DOI: 10.5281/zenodo.24305045

*Abstract— Holistic intelligence emphasizes integrating IQ, emotional quotient (EQ), physical quotient (PQ), and spiritual quotient (SQ) rather than viewing them as separate entities. In the 21st century. The present study assesses the level of holistic Intelligence and the interaction effect among boys' trainee teachers studying in state universities of Haryana. The researcher used a self-made inventory for data collection through a stratified random technique. The results showed that the majority of boys' trainee teachers had a moderate level. There was a significant difference in holistic intelligence among boys' trainee teachers with respect to the universities. A significant interaction effect was shown between locale and university. The study suggests redeveloping the pedagogical practices in teacher training institutes, with the integration of intellectual quotient IQ, emotional quotient (EQ), physical quotient (PQ), and spiritual quotient (SQ). The study also suggests a need for new programs to strengthen holistic Intelligence.*

*Keywords — Holistic intelligence, intellectual quotient, emotional quotient, spiritual quotient, physical quotient, Trainee Teachers.*

## I. INTRODUCTION

### 1.1 Intellectual Quotient (IQ)

Intellectual quotient confers mental abilities like reasoning, conflict management, retrieving information, and evaluating data. IQ plays a remarkable role in achieving academic success and acquiring knowledge, which is crucial for aspiring teachers. However, intellectual quotient alone cannot be sufficient for handling the complexities of strategic classroom management. The term "intelligence" was coined by Stern in 1990.

### 1.2 Emotional Quotient (EQ)

Emotional quotient refers to the ability to recognize, balance, and express emotions effectively. Emotional quotient assists classroom management, strategizes student engagement, problem-solving, and sensitivity to multiple learner needs in teacher education. Indian studies specify gender differences and contextual influences in EQ among trainee teachers.

### 1.3 Spiritual Quotient (SQ)

Spiritual quotient involves self-awareness, meaning formulation, ethical stance, and strategic action. SQ contributes to ethical decision-making, teacher identity formation, and resilience. Spiritual quotient is the sum of Intellectual and Emotional quotient. Indian research magnifies its relevance for teaching effectiveness.

### 1.4 Physical Quotient (PQ)

Physical quotient helps in coping with stress, enduring stamina, and balancing emotions. It demonstrates a commitment to healthy living. Additionally, physical quotient enhances emotional and cognitive functioning, including focus, memory, and efficiency.

## II. LITERATURE REVIEW

### 2.1 Intellectual Quotient among trainee teachers

Bardach et al (2020) studied IQ among prospective teachers, which is less extensive compared with EI and SI. Studies on teacher cognition suggest IQ remains essential for analytical thinking and subject mastery. However, cognitive intelligence alone does not predict efficacy in complex classroom settings.

Mehta et al (2025) studied the triad of EI, IQ, and SI and argued for integrated teacher development models where cognitive, emotional, and spiritual competencies synergistically enhance teaching effectiveness and educational innovation.

### 2.2 Emotional quotient among trainee teachers

Heera & Kaur (2025) studied pre-service teachers and found significant gender differences in Emotional intelligence using standardized tools, with female trainees often reporting higher EI scores than males, indicating the need for gender-responsive training.

Kumar & Sharma (2021) compared prospective and in-service teachers in Delhi NCR and showed the

significance of EI in effective knowledge transfer and teaching performance.

Sharma & Jha (2019) reviewed and emphasized that Emotional Intelligence contributes substantially to teachers' capabilities for self-regulation, empathy, and social awareness — often more so than traditional IQ measures. The studies indicate that emotional intelligence (EI) is crucial for student involvement in the classroom, building interpersonal connections, and fostering professional resilience, which is in line with the wider international emphasis on socio-emotional skills in education.

**2.3 Spiritual Quotient among trainee teachers**

Padhan et al (2025) studied the spiritual intelligence of prospective teachers in Chhattisgarh and Odisha and revealed significant variations based on gender, locality, and academic stream, suggesting SQ is a malleable dimension influenced by demographic factors.

Madhesia & Gaur (2024) found differences in SI among B.Ed. students of Uttar Pradesh from government-aided and self-financed colleges, indicating that institutional context shapes spiritual-cognitive development.

Shobha (2025) studied a correlational study among student teachers in Kerala, which showed significant positive relationships between EI and SI across gender and educational qualifications, underlining the intertwined nature of emotional and spiritual competencies in teacher readiness.

These results showed that spiritual intelligence is strongly linked to values, purpose orientation, and reflective practice—elements essential for comprehensive teacher development.

**2.4 Review on Physical quotient**

Ismail et al. (2020) explored the models of relationship between emotional, spiritual, physical, and social intelligence, resilience, and burnout and revealed that emotional, physical, and spiritual intelligence are positively related to resilience.

## III. DISCUSSION

The researcher explored the level and interaction effect of Holistic Intelligence among trainee teachers, which includes four areas - intellectual, emotional, physical, and spiritual quotient.

The studies support a model where Holistic Intelligence includes intellectual quotient, emotional quotient, spiritual quotient, and physical quotient:

**Dimension wise explanation.**

| Dimensions | |
|---|---|
| Intellectual Quotient (IQ) | serves as the cognitive basis for knowledge acquisition and academic tasks. |
| Emotional Quotient (EQ) | Improves socio-emotional abilities essential for classroom interactions and supporting students. |
| Spiritual Quotient (SQ) | Deepens reflective practices, ethical considerations, and the search for meaning in teaching |
| Physical Quotient (PQ) | Aids in maintaining energy levels, managing stress, and regulating emotions |

**Table -1**

## IV. STATEMENT OF THE PROBLEM

Based on the literature review, the researcher presents the research problem as follows: Assessment of Holistic Intelligence among Boys Trainee Teachers at State Universities of Haryana.

## V. OBJECTIVE

- To study the levels of holistic intelligence
- To explore the main effects of Locale (Rural/Urban)
- To explore the main effects of the University
- To explore the interaction effect on the Holistic Intelligence of trainee teachers.

## VI. HYPOTHESES

- There is no significant variation in holistic intelligence among trainee teachers.
- There is no significant variation in the Holistic Intelligence of trainee teachers with respect to Locale.
- There is no significant difference in the Holistic Intelligence of trainee teachers with respect to the university.
- There is no significant interaction effect between Locale and University on the Holistic Intelligence of trainee teachers.

## VII. SAMPLE

The sample was boys' trainee teachers of state universities from Haryana, India.

## VIII. METHOD OF ANALYSIS

Holistic Intelligence scores of boys were computed to classify respondents into Low, Moderate, and High levels; percentile cut-offs were applied. Mean and standard deviation; two-way ANOVA was used.

**Level of Holistic Intelligence with respect to universities.**

| University | N | Mean | SD | Level |
|---|---|---|---|---|
| Indira Gandhi University, Meerpur–Rewari | 68 | 260.00 | 0.00 | High |
| Chaudhary Bansi Lal University, Bhiwani | 60 | 260.00 | 0.00 | Moderate-High |
| Kurukshetra University, Kurukshetra | 59 | 259.98 | 0.29 | Moderate |
| Maharishi Dayanand University, Rohtak | 57 | 260.00 | 0.00 | Moderate |
| Guru Jambeshwar University, Hisar | 62 | 260.00 | 0.00 | Low-Moderate |

**Table-2**

**Interaction Effect of Holistic Intelligence with respect to locale and universities**

| University | Locale | Level of Holistic Intelligence |
|---|---|---|
| Indira Gandhi University, Meerpur–Rewari | Rural | Moderate |
| | Urban | High |
| Chaudhary Bansi Lal University, Bhiwani | Rural | Moderate |
| | Urban | Moderate-High |
| Kurukshetra University, Kurukshetra | Rural | Moderate |
| | Urban | Moderate |
| Maharishi Dayanand University, Rohtak | Rural | Moderate-High |
| | Urban | Low-Moderate |
| Guru Jambeshwar University, Hisar | Rural | Low |
| | Urban | Moderate |

**Table-3**

## IX. RESULTS

Indira Gandhi University, Meerpur–Rewari found a higher level of holistic intelligence among trainee teachers, while Chaudhary Bansi Lal University, Bhiwani trainee teachers showed a moderate to high level. Trainee teachers of Kurukshetra University and Maharishi Dayanand University showed a moderate level, and Guru Jambheshwar University, Hisar and Chaudhary Devi Lal University, Sirsa showed low to moderate levels. The study showed that trainee teachers from urban and rural areas were significantly different in their holistic intelligence. Urban boys showed higher holistic intelligence scores than rural trainee teachers.

A significant interaction effect was shown between locale and university. The study showed that the influence of locale on holistic intelligence was different with respect to universities. In some universities, rural–urban differences were weak, while in others, the gap was strong. Urban boys consistently exhibit higher levels of holistic intelligence than rural boys across almost all universities. Urban trainee teachers of Indira Gandhi University showed a high level, whereas rural boys remained at the moderate level, indicating the positive influence of urban exposure and institutional resources. Urban trainee teachers of Chaudhary Bansi Lal University and Maharishi Dayanand University showed a moderate–high level compared to rural boys with a moderate level. Urban trainee teachers of Chaudhary Devi Lal University and Kurukshetra University and Guru Jambeshwar University showed moderate level, whereas rural trainee teachers of these universities showed difference among rural trainee teachers showed moderate low level, low–moderate, respectively.

## X. CONCLUSION

This study utilizes holistic intelligence as its overarching theoretical framework to combine Indian research on intellectual, emotional, physical, and spiritual quotient among trainee teachers. By including all four areas, the study showed isolated interpretations and provided a comprehensive understanding of teaching effectiveness. The reviews also show that studies highlight the importance of emotional quotient (EQ), spiritual quotient (SQ), and Physical quotient (PQ) as essential complements to intellectual quotient in preparing future teachers. Incorporating holistic intelligence frameworks in teacher training can improve pedagogical skills, classroom management, and moral values.

## XI. EDUCATIONAL IMPLICATIONS

1. Universities should design holistic development programmes, especially for rural trainee teachers.

2. Best practices with minimal rural–urban gaps should be reproduced across institutions.

## REFERENCES


[1] Padhan, R. K. Behera, C. Padhan, and P. K. Behera, "Spiritual intelligence of prospective teachers: A study," i-manager's Journal on Educational Psychology, vol. 18, no. 4, pp. 78–85, 2025.

[2] P. Heera and A. Kaur, "A study of emotional intelligence of pre-service teachers," International Education and Research Journal, 2025.

[3] K. Madhesia and S. Gaur, "Comparative study of spiritual intelligence of prospective teachers," International Journal of Advance and Applied Research, 2024.

[4] K. U. Mehta and A. Kumar, "Nurturing the whole teacher: The primacy of emotional intelligence in the triad of EQ, IQ, and SQ," International Education and Research Journal, 2025.

[5] P. Heera and A. Kaur, "A study of emotional intelligence of pre-service teachers," International Education and Research Journal, vol. 3, no. 1, pp. 45–48, 2017.

[6] S. Kumar and R. Sharma, "Emotional intelligence and teaching performance among prospective and in-service teachers in the Delhi NCR region," International Journal of Psychosocial Rehabilitation, vol. 25, no. 2, pp. 1123–1131, 2021.

[7] D. Singh and A. Jha, "Emotional intelligence and its role in teacher effectiveness," International Journal of Educational Psychology, vol. 1, no. 1, pp. 45–52, 2012.

[8] N. A. Author and S. Shobaba, "Emotional intelligence and spiritual intelligence: A correlational analysis among student teachers in Malappuram district," International Journal of Environmental Sciences, vol. 11, no. 12s, pp. 539–545, 2025, doi: 10.64252/we5ak526.

[9] L. Bardach and R. M. Klassen, "Smart teachers, successful students? A systematic review of the literature on teachers' cognitive abilities and teacher effectiveness," Educational Research Review, vol. 30, Art. no. 100312, 2020, doi: 10.1016/j.edurev.2020.100312.

[10] K. U. Mehta and A. Kumar, "Nurturing the whole teacher: The primacy of emotional intelligence in the triad of EQ, IQ, and SQ," International Education and Research Journal, vol. 11, no. 1, 2025, doi: 10.5281/zenodo.15592305.

# Cloud-Based Enterprise Information Systems Are Secure Using Zero Trust Architectures.

Mrs. Priyanka Sharma
Assistant Professor, Department of Management, Guru Kashi University, Talwandi Sabo, Bathinda
Email Id: priyankasharma2062026@gmail.com

Accepted: 20-09-2026 Published: 12-10-2026 Page No. 9-13 DOI: 10.5281/zenodo.24330068

*Abstract— Enterprise information systems are now deployed and hosted in a flexible cloud or app-based environment rather than in a fixed, perimeter-protected data centre. In an environment like this, no device or user in the network is ever automatically trusted. In addition, a compromised credential threatens the entire action set of an adversary. In accordance with the principle of "never trust, always verify", this paper proposes zero trust architecture with layered approaches for cloud-based enterprise information system. How does the underlying technical architecture function? The study used a design-science research method. The design requirements were established by literature. We then propose a five-layer structure containing identity, device, network, application and data. The next step is to outline the decision and the flow of enforcement with monitoring.*

*For evaluating design effectiveness six criteria were suggested. Testing is carried on using classical perimeter model. The criteria include limiting the area for attack, avoiding lateral spread, fine-grained access control and others. According to the analysis, the model displays superior containment and visibility. Nonetheless, its complexity at the outset is high. Use staff training programme during first test run administratively required. Extensive SEO-based content writing. A reusable reference design, evaluation framework and prioritization of the control risk model will be available to organizations.*

*Keywords — Zero trust, cloud security, information systems, access control, micro-segmentation, continuous monitoring, design science.*

## I. INTRODUCTION

The IT of today is performing day-to-day office functions like writing email, creating documents, etc. The notion that remote workers are inside and outside the business almost meaningless because of the SaaS applications and cloud services of the business. Cloud computing is a model that allows access to a shared network resource. According to NIST, it is a model that requires. Businesses consider a service a great choice due to its flexibility and cost efficiency, scalability [2].

This creates a problem regarding the security. A conventional firewall is established to deny access and functions as a barrier between a trusted internal network and an untrusted external network. So if an attacker manages to take one identity or device, they have access to many other devices. The attacker can shift across the network. According to the zero trust model, both external and internal networks are not trusted. This makes the workings of this model more stringent and practical. Requests ought to be checked explicitly according to the identity of the requester, the state of the device as well as the context of the request.

An increasing number of organizations and government agencies are adopting the zero trust security models. Countless organizations are unable to generate a useful configuration.

RQ1.What are the requirements on the design of a zero-trust architecture to secure cloud-hosted information systems?

RQ2. In what way can we present or categorize these requirements into multilayer types for implementation purposes?

RQ3. How does the proposed architecture compare to a classic perimeter model in operational and security terms?

The resultant design requirements derived include a five-layer reference architecture with policy decision flow, analytical evaluation framework consisting of six criteria, simple risk model for prioritization of controls and a phased adoption roadmap.

## II. HISTORY & IMPORTANT ASSIGNMENTS.

**A. The design of security models for businesses:**

The phrase "zero trust" was coined by Kindervag who thought we ought to treat all traffic on a network as untrusted, log it, inspect it. As detailed by Google's BeyondCorp, an industry implementation, access decisions depend on user and device, not network location. This means that workers can safely carry out their work without a VPN [5]. Ultimately, these concepts were combined and included in a reference model defined at NIST [3]. Simply put, Zero Trust is a fanciful term that denotes implementation of certain principles by an organization, which helps protect its users, assets, and resources. We will also introduce the policy enforcement point and the policy decision point. According to Gilman and Barth, the practitioner's perspective can build comparable frameworks.

**B. Survey of materials used:**

Syed et al. reviewed literature on zero trust after which they divided it into architecture, technologies and challenges. Note 7 states that actual assessments on application and effectiveness have very few research papers.

**C. Politics of data and accessibility:**

Abnormal behaviour detection is the basis for continuous verification. Authors of the reported study something about using a detection and intrusion detection system of anomaly in the network using processes of machine learning techniques. As Sommer and Paxson explain, it is hard to detect attacks against computer networks. In this respect, they state that misclassification is expensive, detection of true novel attacks is hard, normal traffic is slow-changing.

According to research papers, ABAC can perform authorization based on user, resource and environment attributes in a more fine-grain manner [10].

Earlier studies mainly explain basic principles and set out existing technologies, while not many cite a reference model binding together identity, device, network, application and data controls in one flow which allows for easy comparison of that design against a perimeter model.

## III. RESEARCH PROCESS.

• Hevner and his co-authors [11] offer the precise concept of the design-science approach to information systems research in terms of the design, justification or evaluation of a purposeful artefact against stated criteria. It goes through four stages.

• The perimeter model of cloud ecosystems has a number of flaws.

• The scrutiny can refer to the studies [3], [5] and [7].

• The design requirements have been presented in section IV-A, the issues have been identified.

• The layers and policy flow arrangement is described having been specified under the artifact design.

• The evaluation compares the architecture to the perimeter model on six pre-defined criteria. All of the criteria are rated low, medium, high level on the basis of justification from 100 words (Section V).

• The assessment that is being held has a quality character and also has an analytical character. The ratings are based on the cited literature; whose reasons are more certain than deduction labour. Section VII provides an overview of the effect of this decision.

## IV. SUGGESTED STRUCTURE

**A. Design Specifications.**

The literature was used to derive six requirements.

• R1: Always verify access requests with identity, device health and context.

• R2: the minimum privilege should be provided which is to be limited scope wise and time wise.

• R3: is a center that makes the assumption of breach in order to segment an environment and restrict lateral movement.

• R4: records and logs every access decision made.

• R5: Make sure that the real data is safe, not just the means of transport.

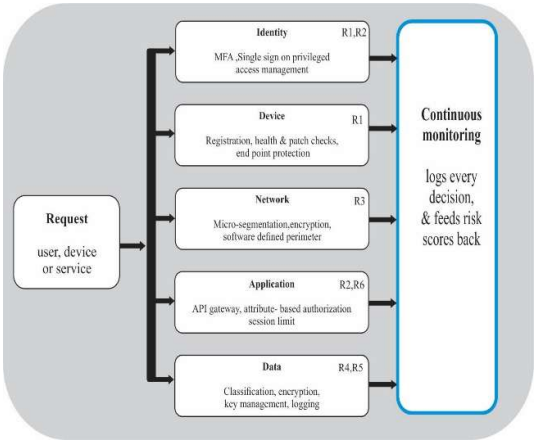
• R6: Stay usable in order that the security controls do not cause users to circumvent them.

**B. Five-Layer Reference Architecture**

| Layer | Main Controls | Requirement Addressed |
|---|---|---|
| Identity | Utilize multiple forms of authentication and single access to login in high privilege access | R1, R2 |
| Device | Device Configuration and Patch Registration for Enhanced Protection | R1. |
| Network | The software-defined perimeters involve network micro-segmentation and encrypted traffic. | R3. |
| Application | The functions of the application API Gateway can use attribute-based authorization, session limits, etc. | R2, R6 |
| Data | The data should be categorized and stored in an encrypted manner at rest and in transit by using appropriate keys which should be responsive to logs | R4, R5 |

*Table I illustrates the levels of a proposed zero-trust architecture.*

**Five layers each verify access and report to one monitoring plane.**




*(Figure 1). A five-layer zero-trust reference architecture with a monitoring plane.*

**C. The Process of Designing and Enacting Policy**. As stated in the NIST model, every access request is intercepted by a policy enforcement point (PEP). It also interrogates a policy decision point (PDP) for a verdict.

A user or service will request access to use the resource to initiate the flow.

• The PEP transfers the identity, device, and context attributes to the PDP for requesting permissions.

• The PDP assesses these criteria with regard to the relevant policy a. The Incident command Centres Risk Score and the latest ABAC rules.

• The PDP may choose to allow, refuse or direct the use of step-up authentication involving an additional authentication component.

• If granted access, the PEP will connect to the resource that has the most minimal privilege level.

• The conduct and verdict are recorded and therefore used on an ongoing basis this can deactivate the session in case of any abnormal behaviour.

**Every request is checked by the PDP before the PEP opens a session.**

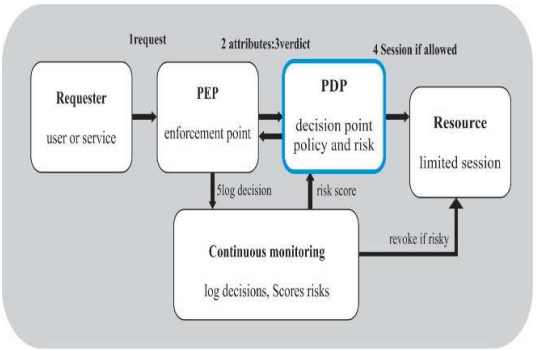



***Figure. 2**. Policy decisions and implementations flow between the PEP, PDP, and monitoring.*

**D. To Prioritize the protection of the various assets**: It is possible to estimate the relevant residual risks, this being achieved by multiplying the probabilities of being compromised by the related impact and then deducted the figures that already account for effective control. The equation $R_i = L_i \times I_i \times (1 - e_i)$ may be used to approximate the asset risk R. $L_i$ refers to the likelihood of compromise in the expression. In contrast, the

I_i indicates the impact, and e_i (which takes value 0-1) gives the overall efficiency of the measures taken. The initial step is to provide assets with a high Li × Ii and a small. The necessity for periodic reviews and updates of the values of L_i, I_i and e_i, based on relevant risk assessment undertaken by the organization, would be essential.

## V. ANALYTICAL EVALUATION.

In table II, there are six criteria to compare the approach against a perimeter model.

| Criterion | Perimeter Model | Proposed Zero Trust Model | Rationale |
|---|---|---|---|
| Attack-surface reduction | Medium | High | Resources are hidden until access is verified |
| Lateral-movement containment | Low | High | Micro-segmentation and per-request checks limit spread |
| Access-control granularity | Low | High | Attribute-based, resource-level decisions |
| Visibility and auditability | Medium | High | Every decision is logged and analyzed |
| Usability | High | Medium | Extra verification may add friction; step-up only on risk |
| Implementation complexity | Low | High | More components, policy design, and integration effort |

***Table 2 provides a qualitative evaluation (Low / Medium / High) of the security models.***

There appear to be alternate equilibria. The proposed model bests perimeter model in the four standards of security as it tests every request and does not assume the trust of the request based on its location in the overall network. Nonetheless, it isn't as user-friendly and is rather complicated to implement. Verification and policy management would require dedicated onboard devices and expertise. The ensuing section explains how cost savings could be achieved through the use of risk-based step-up authentication. On account of the fact that detection takes place through the analytics [9], we can see continuous monitoring as an additional control rather than an alternative to strong identity management and segmentation.

## VI. PLAN FOR EXECUTION.

The organization can choose to slow down the process so that the risk is lower and learning takes place. As demonstrated in Fig. 3, the roadmap consists of four phases which flows through audits and governance.

• Begin with a careful evaluation of the users, devices, programs and data on the environment and messages between them. Make use of the risk model in section IV-D to risk rank the assets.

• The identity is the second phase The obligated implementation of MFA and SSO, as well as a privileged account review.

• Establish frequent evaluations and revisions due to alterations that happen in business practices. When a firm's goals and objectives change; policies of the firm shall also be changed. It is therefore advisable to review the policies monthly. A bet policy is one which is always being improved and never finished.

• The design phase must be monitored in the fourth step of the 5-step approach to incident response planning. Logs are centralized, typically in a security information and event management (SIEM) system. The outcomes of tested incidents are regularly examined for verification.

**Adoption in four phases, with governance running througout**

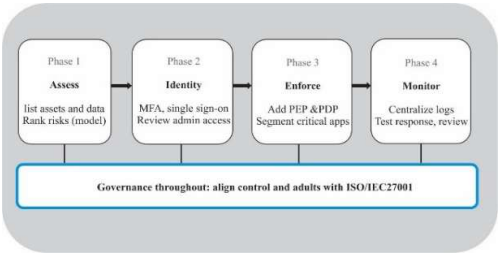



*Figure. 3. A roadmap built of four phases, with continuous governance.*

An Information Security Management System (ISMS) such as ISO/IEC 27001 allows the controlled documentation and repeatable responsibility and audit through all four sizes.

## VII. LIMITATIONS

The review which is the basis for this study suggests that the measure of the table II is not measured, but reasoned. The outcomes will differ as per deployments and several organizations size, fund etc will differ the outcomes. A risk model is a straightforward assessment of the likelihood and impact whose quality depends on the assessment of inputs. Future research should focus on creating a testbed or case-study organization to facilitate metric assessments. The result measures which will cause a better understanding of deploying a zero trust architecture. Also, the objectives are to measure delay in authentication, attempts of lateral movement being blocked and time taken to detect incidents by a bunch of organizations. Additionally, businesses can evaluate user experience and productivity, as well as the relevance of zero trust to multi-cloud, IoT, and OT (operational technology) scenarios through the rollout.

## VIII. CONCLUSION

A zero trust reference architecture for the cloud-based enterprise information systems is made using five layers. For every access request, there is a policy decision flow and enforcement flow. According to our analyses, containment, granularity and visibility model is superior to a perimeter model. The added complexity of the benefits is causing users to experience friction. A systematic priority roadmap employs a basic risk assessment model to ensure that efforts are aimed at major risk activity. The functional and non-functional requirements are not adequately validated.

## REFERENCES


[1] P. Mell and T. Grance, "The NIST definition of cloud computing," Nat. Inst. Standards Technol., Gaithersburg, MD, USA, Spec. Publ. 800-145, Sep. 2011.

[2] M. Armbrust et al., "A view of cloud computing," Commun. ACM, vol. 53, no. 4, pp. 50–58, Apr. 2010.

[3] S. Rose, O. Borchert, S. Mitchell, and S. Connelly, "Zero trust architecture," Nat. Inst. Standards Technol., Gaithersburg, MD, USA, Spec. Publ. 800-207, Aug. 2020.

[4] J. Kindervag, "Build security into your network's DNA: The zero trust network architecture," Forrester Research, Cambridge, MA, USA, Nov. 2010.

[5] R. Ward and B. Beyer, "BeyondCorp: A new approach to enterprise security,"; login, vol. 39, no. 6, pp. 6–11, Dec. 2014.

[6] E. Gilman and D. Barth, Zero Trust Networks: Building Secure Systems in Untrusted Networks. Sebastopol, CA, USA: O'Reilly Media, 2017.

[7] N. F. Syed, S. W. Shah, A. Shaghaghi, A. Anwar, Z. Baig, and R. Doss, "Zero trust architecture (ZTA): A comprehensive survey," IEEE Access, vol. 10, pp. 57143–57179, 2022.

[8] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," IEEE Commun. Surveys Tuts., vol. 18, no. 2, pp. 1153–1176, 2nd Quart., 2016.

[9] R. Sommer and V. Paxson, "Outside the closed world: On using machine learning for network intrusion detection," in Proc. IEEE Symp. Security Privacy, Oakland, CA, USA, 2010, pp. 305–316.

[10] V. C. Hu et al., "Guide to attribute based access control (ABAC) definition and considerations," Nat. Inst. Standards Technol., Gaithersburg, MD, USA, Spec. Publ. 800-162, Jan. 2014.

[11] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design science in information systems research," MIS Quart., vol. 28, no. 1, pp. 75–105, Mar. 2004.

[12] Information Security, Cybersecurity and Privacy Protection — Information Security Management System.

# Effect of Mindfulness Modules On Teaching Effectiveness of Prospective Teachers

[1]Ms. Neema S. Nair, [2]Dr. Harish Mittu
[1]Research Scholar, School of Education, Lovely Professional University, Phagwara,
[2] Professor, School of Education Lovely Professional University Phagwara
Email: [1]neema01phd@gmail.com, [2]harish_mittu@rediffmail.com

Accepted:20-09-2026 Published:12-10-2026 Page No:14-18 DOI: 10.5281/zenodo.24330686

***Abstract***— *The NEP (2020) has established rehabilitation through the formulation of holistic teachers to enable the student to be balanced image. Teacher education is thought to address these requirements through the mindfulness-based approaches. This study assessed a structured mindfulness module's effect on the teaching effectiveness of prospective teachers by comparing control versus experimental groups. A design with a design group test was adopted. Eighty-three B.Ed. students formed the sample. Teachers from a Teacher Education Institution in South Kerala. There were two groups of teachers: an experimental group of 47 teachers who received a 6-week mindfulness module and a control group of 36 teachers who received regular training. Modules for mindfulness, tools for mindfulness pedagogy, lesson plans that are integrated, and both self-constructed and standardised self-evaluative scales for measuring the effectiveness of mindfulness-integrated teaching (Cronbachs α = 0.88). Independent samples z-test was used to analyze data. Students in the experimental group and students in the control group did not differ significantly on the pre-test (t = 0.54 and p = .593 > .05). This is reflected in the means and standard deviations stated as such: 34. The experimental group made significant improvement with t score 14.21 from pre-test to post-test level. The post-test differentiating factor between the experimental and control groups was the mean score where the experimental group scored significantly higher (M = 115.04, SD = 11.67) than the control group (M = 81.78, SD = 17.55) where t = 9.83, p < .01. According to the study, the teaching effectiveness of prospective teachers got enhanced by mindfulness modules. The need for inclusion of mindfulness in teacher education curricula is proved by these findings.*

*Keywords: Mindfulness Modules, Teaching Effectiveness, Prospective Teachers, Teacher Education, and NEP (2020).*

## I. INTRODUCTION

Educators must be trained to efficiently and effectively manage a heterogeneous class. According to NEP 2020, teachers need to be reflective and self-aware individuals in order to teach holistically. Our college, like other alternative teacher training colleges, provides energising rather than draining teacher training programmes which are enthusiastic and emotionally fulfilling.

Teachers usually submit enthusiastic applications for the use of mindfulness in anticipation of the issues that will occur. Mindfulness is being present in an accepting way. Self-regulation, classroom management, and situation response are essential teachers' competencies (Kutube-Zinn, 1994). Mindfulness is valued in India. Focus is achieved by getting equanimous using the teachings from the Bhagavad Gita and restraining the restless mind. Mindfulness is derived from the word smriti in Sanskrit, and from onerata in Pali. Mindfulness has been found, in various studies, to increase one's focus, ability to regulate emotion and stress relief (Bishop et al., 2004) This mindfulness supports better teaching competence and lessens educational burnout. Research suggests that mindfulness training boosts the confidence and teaching skills of early-career instructors. Currently, there have not been any structured mindfulness modules for Indian teacher education (Walsh 2017; Garner 2018). The purpose of the present paper is to fill this gap by assessing the impact of a mindfulness module.

## II. LITERATURE REVIEW

According to Bai (2001), having knowledge gained from education does not prompt them to take an action in future events.

Staying present helps to avoid a gap. As per Weisbaum (2014) and Hoyt (2009), being mindful can reduce classroom stress. Research with Future Educators Yields Positive Results. According to Garner (2018), teachers' wellbeing improved and training strategies used in class were enhanced after six-week mindfulness-based social emotional learning (SEL) training. Instances of intentional teaching occurred more often. A study by Roeser and 2007 and Kerr 2017 revealed that resolving psychosocial problems can enhance the engagement of students in learning.

There is greater self-awareness and self-regulation of emotions. A study conducted in India by Ashwathi and Kalaimathi found that the prospective teachers have low-level baseline mindfulness. We need to intervene because of this. As the literature suggests, we need to devise modules linking mindfulness with teacher readiness. The present study of the researcher aims at developing a prototype module and testing it. In addition, the investigator also tries to investigate the impact of module on the basis of gender, locale and pedagogy subject.

## III. OBJECTIVES

The study pursued the following objectives:

1. To assess competitive anxiety among male and female university rugby players using the SCAT;
2. To compare competitive anxiety between male and female players; and
3. To examine the relationship between competitive anxiety and catching, tackling, and passing performance.

The rationale was that rugby combines technical execution with intense physical contact and time pressure. Understanding whether anxiety is associated with skill performance may assist coaches and sport psychologists in deciding when anxiety-management training is warranted [1], [2].

## IV. METHODOLOGY

The study used a Quasi-experimental pre-test and post-test type. The study's sample was made up of 83 B.Ed. students. Prospective teachers from one Teacher Education Institution in South Kerala. The experimental group included 47 pre-service teachers, while the control group had 36 pre-service teachers. Mindfulness modules were experimented with to estimate their effective impact on the teaching effectiveness of prospective teachers. The selection of the prospective teachers was done on the basis of their level of teaching effectiveness using a standardized scale teaching effectiveness for prospective teachers (Nagpal & Kumar, 2022). Training was given through the mindfulness module to the experimental group. To collect more quantitative data, the study uses quasi-experimental methods in the form of a pre-test and post-test research design. The prospective teachers received the modules over a period of 6-weeks and from the third module onward, the trainees were given assignments to develop and demonstrate mindfulness pedagogic tools. In their lesson preparations, the teachers incorporated unique activities as well as pedagogic tools to make conceptual understanding easier for students. In the sixth session, the lesson plans were corrected further and finalised for effective integration of mindfulness pedagogic tools.

Moreover, classroom observations were conducted to ascertain the implementation of the subject-wise lesson plans, the use and integration of Mindfulness Pedagogic Tools (MPT), feedback, related activities, extended activities, emotional attunement of teachers, class management and effective use of time in teaching of for the prospective teachers. After every observation, a remedial feedback with specific inputs was provided to enhance the perspective's teaching effectiveness. The prospective teachers in the control group received regular teacher education. They did not go through mindfulness training.

Tools: The researcher constructed a mindfulness-integrated teaching effectiveness scale to assess the effectiveness of the mindfulness modules delivered by teachers. The researcher used self-evaluative and standardized mindfulness-integrated teaching effectiveness scale in the present study. Nair, Neema S., and Dr. Harish Mittu. Mindfulness-integrated Teaching Effectiveness Scale. University of the West of England, 2026.

## V. RESULTS

This study aims to determine whether the undocumented the experimental group and the control group prospective teachers differ significantly in terms of the pre-teaching effectiveness. Teaching effectiveness of 47 experimental and 36 control group prospective teachers was measured before the application of the mindfulness-based modules. A z-test for independent samples was used to analyze the significance difference in pre-teaching effectiveness of experimental group and control group. Teachers effectiveness ratings of prospective teachers were used to collect data. The pre-test averages and standard deviations were

computed. The experimental group and control group prospective teachers have similar pre-teaching effectiveness claims the Ho hypothesis.

***Table 1 Significance of the Difference in Means of Pre-Teaching Effectiveness Scores of Experimental Group and Control Group Prospective Teachers.***

| Group | N | Mean | SD | $SE_D$ | z-value | p-value | Remarks |
|---|---|---|---|---|---|---|---|
| Experiment | 47 | 78.15 | 13.06 | 3.44 | .54 | .593 | $H_0$ Accepted |
| Control | 36 | 80.00 | 17.21 | | | | |

As shown in Table 1, the experimental group's mean score was 78.15 (SD = 13.06), reflecting pre-teaching effectiveness. Similarly, the control group's mean score of 80.00 (SD = 17.21) also indicated pre-teaching effectiveness. Being the SED (3.44), the above values are not significantly different. The z-value obtained as 0.54 (p = 0.593,> .05) is statistically insignificant indicating an absence of any significant difference in the pre-teaching effectiveness between the experimental group and control group, prospective teachers.

## VI. EXPLANATION

This examination was useful to verify the initial comparability of both groups and to.

It is vital to make sure that the differences will only be due to the mindfulness intervention during the post-test stage. It was found that the mean score of teaching effectiveness of the experimental group was 78.15 (SD = 13.06) and that of control group 80.00 (SD = 17.21). The z-value of 0.54 with p-value of 0.593 was not statistically significant at 0.05 level. The findings revealed no significant difference between the experimental and control groups regarding teaching effectiveness prior to the implementation of the mindfulness module. As a result, the null hypothesis which states that there is no significant difference in pre-teaching effectiveness of experimental group and control group prospective teachers was accepted. The lack of statistically significant difference reveals that both groups had similar teaching effectiveness before the mindfulness modules were administered.

Thus, it seems a valid population to assess the impact of the mindfulness modules during the experimental phase. Because of the near-equal starting levels of teaching effectiveness of both groups, the observed differences in the post-test stage may be attributed to the mindfulness intervention. As a result, the pre-test analysis successfully established the groups' baseline equivalence. This equivalence enhances the study's internal validity, further strengthening the evaluation of the effectiveness of the mindfulness modules on teaching effectiveness in the post-test.

Result Pertaining to Difference in Post-Teaching Effectiveness of Experimental Group and Control Group Prospective Teachers

To examine whether there is a significant difference in post-teaching effectiveness of experimental group and control group prospective teachers, teaching effectiveness of 47 experimental and 36 control group prospective teachers was measured after the implementation of the mindfulness-based modules. Independent samples z-test was applied to study the significant difference in post-teaching effectiveness of experimental group and control group prospective teachers. The data were collected in terms of rating scores of prospective teachers on teaching effectiveness. The means and standard deviations of post-tests were computed

Ho: There is no significant difference in post-teaching effectiveness of experimental group and control group prospective teachers.

**Table 2 Significance of the Difference in Means of Post-Teaching Effectiveness Scores of Experimental Group and Control Group Prospective Teachers.**

| Group | N | Mean | SD | SED | z-value | p-value | Remarks |
|---|---|---|---|---|---|---|---|
| Experiment | 47 | 115.04 | 11.67 | 3.39 | 9.83 | .000 | $H_0$ Rejected |
| Control | 36 | 81.78 | 17.55 | | | | |

## VII. INTERPRETATION

According to table 2, the mean score for the experiment group was 115.04 (SD = 11.67), while the mean for the control group was 81.78 (SD = 17.55). The SED is 3.39 which shows precision of the paired comparison. The calculated z-value of 9.83 (p = 0.000 > 0.01) has proved to be statistically significant which confirms that there is significant

difference in teaching effectiveness of experimental group and control group prospective teachers.

## VIII. DISCUSSION

The experimental group had a mean score of 115.04 (SD = 11.67) and control group mean score was 81.78 (SD = 17.55). The z-value obtained (9.83) proved to be highly significant ($p < .001$). It indicates that prospective teachers provided mindfulness modules were more effective in teaching than prospective teachers in control group. The average difference that is visible means the mindfulness modules training had a significant positive influence on the prospective teachers' teaching effectiveness. The post-test difference can be attributed to the mindfulness modules given that pre-test analysis confirms that teaching effectiveness between both groups is statistically equivalent. Existing literature on mindfulness and teacher effectiveness provides strong support for the present findings. According to Jennings et al. (2011), mindfulness training enhances teachers' capacities to respond optimally to classroom challenges, regulate their emotions, create supportive classroom environments, and interact with students in a caring manner. The significant progress of the experimental group may be accounted due to the mindfulness modules which may have led to a better teaching process. The results align with the findings of Baptiste (2014), Walsh (2017), Garner (2018) that there was an increase in self-attunement, emotional regulation as well as a better awareness of the emotional and attentional needs of students through mindfulness-based teacher training. The enhancements help teachers plan their classroom situations in a more focused, reflexive and adaptive manner to teach well. The scores of teaching effectiveness are high showing that they are effective. The experimental group in the current study may demonstrate these improved professional skills. Furthermore, Roeser et al.'s (2013) and Kerr et al.'s (2017) works together with those of Zimmerman (2018), Park et al. (2020) and Hirshberg et al. (2020) found that employing mindfulness helped pre-service teachers improve their resilience, self-confidence, self-awareness, empathy and occupational acceptance. The personal and professional traits that enhance teaching practices of teachers may also explain the better performance of the experimental group. According to Kaur (2024) and Chen (2024), pre-service teachers exhibit a positive relationship between mindfulness, cognitive maturity and cognitive flexibility.

The study results suggest that mindfulness training not only reduces stress but also aids in professional development. This can improve teaching skill, classroom management, emotional regulation, interpersonal sensitivity and reflexivity or reflection on one's own practice. The experimental group showed significantly more improvement than the control group, confirming the effectiveness of mindfulness modules in enhancing the teaching effectiveness of prospective teachers.

## IX. CONCLUSIONS

Courses on mindfulness can make a teacher effective. An enhancement in the efficacy of teaching pertains to improved teacher awareness, acceptance, curiosity and compassion which greatly contribute to the improvement in reflective teaching, teacher resilience and teacher student relationships. The prospective teachers in the experimental group produced outcomes that were significantly better than those of their control group counterparts, which had no orientation. Incorporating mindfulness into teacher education manifold will prepare reflective, competent and emotionally balanced teachers for the class of 21 century ones. Mindfulness Pedagogic Tools can be integrated into lesson plans in the teaching-learning process of prospective teachers to transform rote learning to reflective learning which will enrich the teaching-learning process. This suggested that collaboration between teacher education institutions and schools may close the gap between theory and practice in teacher education. According to the Impact of mindfulness study, putting mindfulness into practice is a professional development that not only reduces stress. Mindfulness implementation enhances teacher performance and management, affect regulation, interpersonal sensitivity, and reflection on teaching.

## REFERENCES


[1].Ashwathi, K., & Kalaimathi, S. (2022). Mindfulness status among prospective teachers. International Journal of Educational Research, 12(2), 45-52.

[2].Bai, H. (2001). Beyond the educated mind: Towards a pedagogy of mindfulness. Journal of Educational Thought, 35(2), 139-153.

[3].Garner, P. W. (2018). Mindfulness-based SEL programming to increase pre-service teachers' mindfulness and emotional competence. Journal of Research in Childhood Education, 32(4), 475-489.

[4].Hirschberg, M. (2020). Integrating mindfulness into pre-service teacher education. Early Childhood Education Journal, 48(3), 321-330.
[5].Hoyt, L. (2009). Teaching with mindfulness: Pedagogy of being-with/for. Educational Theory, 59(3), 259-278.
[6].Kerr, S. L., et al. (2017). Impact of mindfulness on self-compassion in pre-service teachers. Teacher Education Quarterly, 44(2), 67-84.
[7].Kumar, D. and Nagpal, S. (2021). Effectiveness of constructivist blended instructional paradigm. Ph.D. Thesis, Lovely Professional University, Punjab, India. Retrieved from https://shodhganga.inflibnet.ac.in/handle/10603/330590.
[8].Kwon, Y. (2015). Benefits of mindfulness to teacher professional practice. Journal of Educational Research, 108(4), 301-309.
[9].National Education Policy. (2020). Ministry of Education, Government of India.
[10]. Roeser, R. W., et al. (2013). Mindfulness training and qualities of teacher mindfulness. Developmental Psychology, 49(2), 222-233.
[11]. Walsh, K. (2017). Effects of mindfulness-based teacher training on developing teachers. Reflective Practice, 18(5), 641-653.
[12]. Weisbaum, E. (2014). Why mindfulness: Examining an emerging educational pedagogy. Canadian Journal of Education, 37(2), 1-23.

# The Efficacy of Deep Belly Breathing (Adhama Pranayama) on Chronic Low Back Pain (CLBP): An Analytical Review

[1] Sandeep, [2] Sakshi Nagru

[1]Research Scholar, Department of Physical Education, Guru Kashi University, Talwandi Sabo, Bathinda.

[2]Assistant Professor, Department of Physical Education, Guru Kashi University, Talwandi Sabo, Bathinda.

Email Id: [1]dr.mohitmonga@gmail.com, [2]sakshinagru@gmail.com

Accepted: 20-09-2026 Published:12-10-2026 Page No. 19 – 21 DOI: 10.5281/zenodo.24332854

*Abstract— Chronic low back pain has become an international risk factor due to the significant disability and economic burden. Due to some limitations and side effects concerning drugs therapy, various mind-body interventions such as Adhama Pranayama are trending in the world today. The phrase Complementary and alternative medicine (CAM) refers to a diverse group of medical and health care systems. Which are not considered as part of a country's own tradition or culture. The current study examined how an 8-week deep breathing intervention added to a standard treatment could help control Chronic low back pain due to pathological mechanisms that have just been discussed, there was a significant decrease of the subjective pain and the improvement of the lumbar range of motion of children which is similar to the improving autonomic balance of the patient which has become the common pattern of the quest. The Sanskrit word "pranayama" is a combination of "prana" and Pranayama essentially means the expansion of your life force /regulation/control [2]. A simple, affordable, and safe practice for chronic low-back pain is Adhama Pranayama. An approach that strives to address the psychological and physical issues of CLBP*

*Keywords: Chronic low back pain treatment using pranayama, body stabilization, and breath awareness for pain relief.*

## I. INTRODUCTION

Chronic low back pain is defined as pain found between the lower margins of the ribcage and the folds of the gluteal muscles which persists for longer than 12 weeks. According to WHO, depression is estimated to be the foremost reason for disability around the world 2020. In 2020, there were 619 million cases of disability from LBP. Many common pathophysiological mechanisms are shared between low-back pain (LBP) and most neurodegenerative disorders including multiple sclerosis, Alzheimer disease, Parkinson disease and amyotrophic lateral sclerosis.

Traditional therapy focuses on structural deficiencies. Approximately eight to nine out of 10 cases are classified as non-specific. In other words, the causes of this condition cannot be isolated. Many adults suffering from severe back pain are aged between the ages of 28 to 42 years. One of the main features back pain is subjective, one of its primary characteristics, People who have experienced low back pain before are likely to experience it again. Often, symptoms of low back pain clear up on their own without treatment. In some cases of LBP, some complaints of LBP do not remit. The onset of any pain is due to a multifactorial development derived from a bio-psycho-social makeup that draws considerable interest. That being the pain can have an overall objective impact on the person's functionality that ought to be documented. Low back pain (LBP) is a huge contributor to disability for LBP persons according to WHO (world-health-organization). Disability means an activity limitation, i.e., as in daily life [8-10]. Moreover, it has an objective measurement. Due to the biopsychosocial changes that people with lower back pain undergo, their relationship with the world may change. The models' limitations are somewhat dealt with in [6, 11-13].

A small number of people link breath and spine stability. Deep belly breathing is a type of breathing technique. Your diaphragm, an organ, allows your body to inhale and aids in stabilizing your lumbar

spine. This study investigates the efficacy of 8 weeks of Adhama Pranayama versus only physiotherapy in relieving pain and pain perception through activity of ANS and functional mobility in patients with CLBP.

## II. OBJECTIVES AND AIMS

The aim of the present research is to find out if deep belly breathing can improve pain control and lumbar mobility.

- Pain relief assessment is done using different scales VAS.
- Enhancement in functionality is the ODI, RMDQ.
- Take a closer look at involuntary modifications in breath charge and HRV.

The clinical pain score and functional disability index will be highly statistically significant difference between two groups. The central brain muscles contract more, while vagal tone increases in the gut. Exaggerated pain reduction. We will collect the data using questionnaires and interviews. Also, we will evaluate the 60 samples with physical and imaging checkups. For eight weeks, participants will undertake an organized pranayama programme. Assessment of physiological measures of pain and mobility pre-and post-intervention. The controls received standard therapy. Data Analysis to Identify Differences between Groups. Selection of Participants. A clinical study will include individuals 30 to 60 years of age, suffering from back pain for more than 3 months. Patients with back pain but without any identifiable organic lesion. Occupational posture does not cause the patients to develop fault. Individuals who have suffered an infection of the spinal cord or abdomen and are afflicted with acute pain for less than three months. design of research. Each group received half of the subjects in total. Control Group A with n=30 received standard physical therapy which included stretching and strengthening exercise as supervised. Group B, which we shall now call the intervention group, almost had a mixture of thirty. The SPT group had 20 minutes of deep belly breathing every day. As to deepen abdominal breathing which referred to adhoma pranayama using a sitting posture which was Sukhasana, Padmasana, etc. which help expands the diaphragm.

## III. RESULTS OF THE OUTCOMES.

- The Visual Analog Scale (VAS) is a measure 0 – 10 for the intensity of pain perception.
- The Oswestry index measures the impact of back pain on managing life.
- HRV is the technique which helps in profiling the activity of the autonomic nerve system and the measurement of stress.

Group B monitored over 8 weeks has shown significant physiological and behavioural differences as compared to the control. Research shows that use pranayama, lessen severity of pain, independence and psychosocial well-being. Physiological markers illustrate an increase in Hb levels and a drop in Cortisol levels. It becomes very simple to practice, thus compliance from patients is high.

EMG analysis in group B reveals the impact of "internal bracing," which indicates better cooperation between the transversus abdominis and multifidus. The lumbosacral motor unit was recruited in a coordinated manner with Adhama Pranayama.

Following the autonomic re-calibration, Group B showed an increase in high-frequency HRV. A shift from sympathetic dominance to parasympathetic dominance leads to an increase in nociceptive threshold and a decrease in the perception of subjective pain.

The TSK scores from the psychological assessments indicate that the intervention group was significantly lower in kinesiophobia than the control group. The way of breathing told the nervous system that it is “safe” which drastically lowers the unconscious guarding of muscles and fear of movement.

## IV. DISCUSSION.

The combination of the structural stability along with neurophysiological pain regulation in Pranayama would be the ideal intervention for rehab. The act of deep belly breathing takes into consideration the CLBP’s different dimensions This might aid in spine stabilization, lowering the stress response, and providing more self-help. The small sample size and non-standardised protocols set limitations. There are two different ways this exercise works.

**Biomechanical Stabilization Reworded.**

The diaphragm is important for proper posture. People with chronic low back pain (CLBP) often restrict the mobility of their diaphragm which can cause it to lock in a high and flattened position. The lumbar spine is deprived of any internal support by this. Take a deep breath from your belly so your diaphragm moves down. This activity elevates the Intra-Abdominal Pressure (IAP). One could think that the building of this pressure comes about “with the pelvic floor, an internal pneumatic cylinder

which off-loads the mechanical loading of the intervertebral discs and stabilises the lumbo-pelvic-hip complex.

***Modulation Influences Both Biochemical and Neurological Responses***

The constant agony emerges because of inflammation and more sensitivity of the nervous system. The vagus nerve will be stimulated with slow deep breathing, which will dampen cortisol and pro-inflammatory cytokines (IL-6) and improve GABA. This biochemical change happens in the somatosensory cortex of the brain, which makes it less sensitive to pain by "lowering the volume."

## V. CONCLUSION AND RECOMMENDATIONS

Deep Belly Breathing is a simple, inexpensive and effective therapy for Chronic Low Back Pain. This instrument works towards the physiological and mechanical imbalance that causes the pain cycle. Whether in conventional therapy or not, it is the addition of diaphragmatic breathing that augments the nervous system reactivity and core motor control. Adhama pranayama is an effective, non-drug treatment for CLBP. Patients can receive medical care in the comfort of their homes. The impact of randomized studies on clinical guidelines should be greater in order to establish results.

## VI. CLINICAL SUGGESTIONS.

- In physiotherapy and rehabilitation cardio, pranayama can be included in clinical protocols. Teach patients self-management for the long term.
- During training phase 1, patients are constantly encouraged to learn to be in a supine position so that they can work on diaphragmatic breathing to clear the diaphragm.
- During the integration phase, breathing breaks of 5 minutes must be introduced in the activity which generally triggers lumbar pain.
- Dynamic sustainability breath work first engages the gentle, active spinal mobility (cat-cow) breath and then movement to desensitize movement fear and support recovery of function.

## REFERENCES


[1]. The World Health Organization - low back pain fact sheet. Geneva: 2023. Management of Chronic Primary Low Back Pain: Non-Surgical Guidelines Approach of traditional medicine.

[2]. The World Health Organization - low back pain fact sheet. Geneva: 2023. Management of Chronic Primary Low Back Pain: Non-Surgical Guidelines Approach of traditional medicine.

[3]. Apkarian A.V., Scholz J. Chronic pain and neurodegenerative disease share common mechanisms. Drug Discovery Today's discount. Machine 2006; 3:319–326. doi: 10.1016/j.ddmec.2006.09.006.

[4]. [3]. Hod D, March L, Brooks P, et al. The worldwide impact of lower back discomfort: Projections from the 2010 Global Burden of Disease research. Ann Rheum Dis 2014;73(6):968-974.

[5]. Knezevic N. N., Candido K. D., Vlaeyen J. W. S., Zundert J. V., Cohen S. P. Pain in lower back. Lancet 2021;398(10294):78–92.

[6]. da Silva T Mills K Brown BT and others The chance of low back pain coming back: A systematic review. Journal of Orthopaedic & Sports Physical Therapy.

[7]. Author Marras WS The multifaceted back: The system of pathways that cause low-back problems. Human Factors and Ergonomics Society 2012;54(6):881–9.

[8]. James SL, Abate D, Abate KH et al. A systematic analysis for the Global Burden of Disease Study 2017 of global, regional and national incidence, prevalence and years lived with disability for 354 diseases and injuries in 195 countries and territories, 1990–2017.

[9]. Lygren H, Strand LI, Anderson B, Magnussen LH. Are the activity limitations self-reported by patients due to back problems represented in ICF core sets for low back pain? ICF core sets for lower back pain. Physiotherapy Res Int, 2014.

[10]. Kienbacher T, Fehrmann E, Tuechler K et al. Alterations in International Classifications of Functioning, Disability, and Health Domains 'Activity/Participation' Anticipated via Patient Reported Outcomes with Global Back Pain Rehabilitation. Continual Journal Pain 2021;37(11):812–9.

# Perceived Benefits of Participation in Yogic Exercises between Rural and Urban University Students: A Comparative Study

**[1] Rajveer Kaur, [2] Sakshi Nagru**

[1] Research Scholar, Department of Physical Education, Guru Kashi University, Talwandi Sabo

[2]Assistant Professor, Department of Physical Education, Guru Kashi University, Talwandi Sabo

Email Id: [2]sakshinagru@gmail.com

Accepted:20-09-2026 Published:12-10-2026 Page No. 22– 25 DOI: 10.5281/zenodo.24334100

*Abstract—This study was conducted to explore perceived benefits of yoga exercise participation based between rural and urban university students enrolled at Guru Kashi University and residing in District Bathinda, Punjab. The total number of participants were forty (n=40), 20 from rural and 20 from urban areas. The purposive sampling technique was used for the current study. The study employed quantitative approach to examine the perceived benefits of yogic-exercise participation between university students with the use of "Exercise Benefits/Barriers Scale (EBBS)". Independent Sample t-test was applied to check the significant differences. All the 40 participants successfully responded the questionnaire and the results of inferential statistics showed that urban students were better in perceiving benefits of yoga exercise participation, t (38) = 16.88, sig = 0.001 (p < 0.05). This study tried to determine how student's area of residence could influence their perceptions about yoga exercise. The result of this study showed that urban students were significantly different than rural students in their perceived benefits of yoga exercise participation. Hence, it was found that perception of yoga exercise participation for university students, residence is a vital component to consider.*

*Keywords: Perceived benefits, yoga exercise, purposive sampling and Exercise Benefits/Barriers Scale.*

## I. INTRODUCTION

Exercise is really good for people. Doing exercise on daily basis has a lot of benefits. Research shows that people who exercise are less likely to get sick with things like heart disease, type 2 diabetes, some types of cancer and osteoporosis. This is what researchers like Chiuve and others found out in 2008 and also Jeon and others in 2007. Krall and Dawson Hughes in 1994 and Kramer and Wells in 1996 and Powell and others in 1987. Exercise can also make people feel better think clearly and be less depressed and anxious. This is what Berger and Motl said in 2000. Colcombe and Kramer in 2003 and Craft and Landers in 1998 and Petruzzello and others in 1991 and Rejeski and Mihalko in 2001.

In fact, Yoga is a type of exercise that's very old. It combines things like poses, breathing and meditation. Yoga is different from types of exercise. One big difference is that yoga is not competitive. Mostly exercises are competitive in nature. Yoga is more about how you feel about yourself not about comparing yourself to others. This is what Marshall said in 1978 and Iyengar in 2001. Other types of exercise often have strong repeated movements and heavy breathing. These exercises can be good for your bones and muscles. They can also make you tired and stressed. Iyengar said this in 2001. Saraswati said it in 1987. These exercises can give you energy but they can also wear out your body and make you feel sick. Iyengar said this too.

Furthermore, Yoga as an exercise is special because it is not like exercises. Exercise like yoga can help people in ways. Yoga as an activity is good for people because it is gentle and non-competitive. People who do yoga say it helps them feel better and think clearly. Yoga is a type of exercise that people should try because it has benefits. Exercise, like yoga is good, for people because it can help them feel better and live lives.

**Table 1 Descriptive statistics of perceived benefits score.**

| Group | N | Mean | SD |
|---|---|---|---|
| **Urban** | 20 | 97.05 | 2.87 |
| **Rural** | 20 | 82.55 | 2.56 |

The table 3.1, presents the descriptive statistics of perceived benefits score of mean and standard deviation of the participation in yogic exercises between rural and urban university students. Urban students demonstrated higher mean scores that rural

students, indicating greater perceived benefits of participation in yogic exercises.

**Table.2 Independent t-test for perceptions towards yoga participation among urban and rural students.**

| Group | Mean | SD | N | T-value | p-value |
|---|---|---|---|---|---|
| **Urban** | 97.05 | 2.87 | 20 | 16.88 | <0.001 |
| **Rural** | 82.55 | 2.56 | 20 | | |

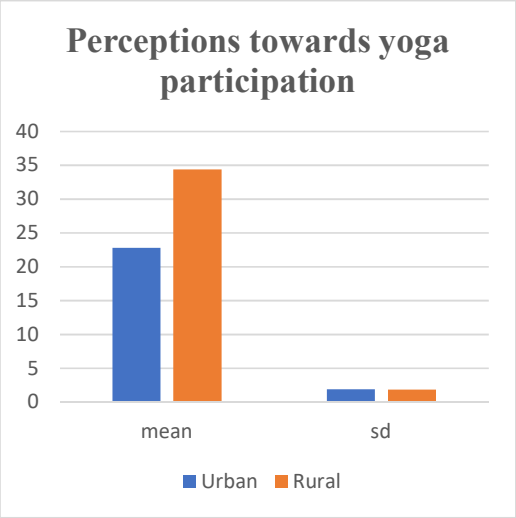
Table 3.2, shows the comparison of perceptions towards yoga participation between rural and urban students. The mean score of urban students was seen to be 97.05 and standard deviation 2.87. the mean score of rural students was 82.55 and standard deviation 2.56. The higher mean score of urban students indicates that they have possessed strong positive perceptions towards yoga participations compared to their rural counterparts.

The obtained t-value of 16.88 was statistically significant at the 0.05 level of significance ($p<0.001$). Since the p-value was less than the prescribed level of significance, the null hypothesis stating that "there is no significant difference in perception of yoga between rural and urban students" was rejected.

The findings suggest that area of resistance significantly influences student's perceptions towards yoga participation. Urban students may have greater exposure to health awareness programs, yoga facilities, trained instructors, and media information related to yoga, which could contribute to their more favourable perceptions. In contrast, rural students may have comparatively fewer opportunities and resources related to yoga participation, resulting in relatively lower perception scores.

Therefore, it was concluded that urban students demonstrate significantly more positive perception toward yoga participation than rural students.

*The bar chart depicts the mean differences of perceptions towards yoga participation between urban and rural students across the groups.*




## II. CONCLUDING REMARKS

The present study aimed to evaluate "Perceptions of benefits of Yoga Participations between rural and urban population" among 40 participants chosen through purposive sampling method, enrolled in Guru Kashi University, Bathinda-Punjab. All participants underwent a structured data collection procedure through a standardized questionnaire "Exercise Benefits/Barriers Scale (EBBS)". After data collection the results has shown significant difference for the selected research objective, mentioned in - Table 2.

## III. DISCUSSION

The aim of the present study was to compare perceptions towards benefits of yoga participation between rural and urban students enrolled in Guru Kashi University and residents of District Bathinda Punjab. Using the adapted Exercise Benefits/Barriers Scale (EBBS), data were gathered and scores were subjected to descriptive statistics and Independent Samples t-test, findings were presented as they relate to study objectives and hypotheses.

## IV. DISCUSSION RELATED TO SELECTED OBJECTIVE

"To assess the perceived benefits of participation in yogic exercises among university students".

Descriptive statistics shown in Table 2, indicated that urban students scored significantly higher (Mean = 97.05, SD=2.87) than rural students (Mean

= 82.55, SD ==2.56) on perceptions toward yoga participation The calculated t-value (t=16.88) was statistically significant at p <0.001). Rejecting the hypothesis about non-significant differences in perceived benefits of yoga participation between rural and urban students.

Urban students think yoga is a thing more than rural students do. This is because urban students see things about health and yoga where they live. They see things like yoga centers and fitness places. They here about the good thing's yoga can do for their bodies and minds. Urban students who are doing yoga and they hear about how it can help them feel better and be healthier. This is why urban students, like the idea of doing yoga than rural students do.

The study I am looking at shows that people who have access to education and healthcare are more likely to know about yoga and have a positive view of it. This is what Mishra and his team found out in 2018. They saw people who know more about yoga and think it is a thing are usually the ones who have better access to education and healthcare.

Similarly, Anand and his team found out in 2023 that people who live in cities are more into yoga and mindfulness than people who live in areas. They saw that city people do yoga and mindfulness often and are more open to it. Other studies, like the ones done by Park in 2026 and Ross in 2013 also show that people who think yoga is good for their body and mind are more likely to do it. So, what this study suggested that residence is a big factor in taking part in yoga exercises as the mindset builds according to the environment, as per the opportunities provided so for. The study is talking about university students. How where they live affect what they think about doing yoga. Yoga participation is what we are looking at here. They study says that where you live matters when it comes to yoga participation, among university students.

## V. CONCLUSION

This study looked at how university students in urban areas think about the good things that come from doing yoga. Researcher found out that there is a significant difference between these two groups of students. This means that where a student lives has an effect on what they think about the things that come from doing yoga. Students who live in cities said they got good things from doing yoga than students who live in the rural areas of country. So, it would be an idea to tell students from rural areas more about the good things that come from doing yoga. This could get more of these students to do yoga and get involved in it. The benefits of yogic exercises are important. We should try to get students, from rural areas to see this and to get more students to participate in yogic exercises.

## REFERENCES


[1]. Sechrist, K. R., Walker, S. N., & Pender, N. J. (1987). *Psychometric evaluation of the Exercise Benefits/Barriers Scale*. *Research in Nursing & Health, 10*(6), 357–365.

[2]. Khalsa, S. B. S. (2004). *Yoga as a therapeutic intervention: A bibliometric analysis of published research studies*. *Indian Journal of Physiology and Pharmacology, 48*(3), 269–285.

[3]. Cowen, V. S., & Adams, T. B. (2005). *Physical and perceptual benefits of yoga asana practice: Results of a pilot study*. *Journal of Bodywork and Movement Therapies, 9*(3), 211–219. https://doi.org/10.1016/j.jbmt.2004.08.001

[4]. Sharma, M., Gupta, N., & Bijlani, R. L. (2008). *Effect of yoga on mental health and quality of life*. *Journal of Alternative and Complementary Medicine, 14*(7), 807–813.

[5]. Ross, A., & Thomas, S. (2010). *The health benefits of yoga and exercise: A review of comparison studies*. *Alternative Therapies in Health and Medicine, 16*(3), 12–20.

[6]. Field, T. (2011). *Yoga clinical research review*. *Complementary Therapies in Clinical*

*Practice, 17*(1), 1–8. https://doi.org/10.1016/j.ctcp.2010.09.007

[7]. Woodyard, C. (2011). *Exploring the therapeutic effects of yoga and its ability to increase quality of life*. *International Journal of Yoga, 4*(2), 49–54. https://doi.org/10.4103/0973-6131.85485

[8]. Büssing, A., Michalsen, A., Khalsa, S. B. S., Telles, S., & Sherman, K. J. (2012). *Effects of yoga on mental and physical health: A short summary of reviews*. *Evidence-Based Complementary and Alternative Medicine, 2012*, Article 165410. https://doi.org/10.1155/2012/165410

[9]. Innes, K. E., & Selfe, T. K. (2012). *Yoga for adults with chronic health conditions: A review of evidence*. *Evidence-Based Complementary and Alternative Medicine, 2012*, Article 960549. https://doi.org/10.1155/2012/960549

[10]. Sherman, K. J. (2012). *Guidelines for developing yoga interventions for randomized trials*. *Evidence-Based Complementary and Alternative Medicine, 2012*, Article 143271. https://doi.org/10.1155/2012/143271

[11]. Cramer, H., Lauche, R., Langhorst, J., & Dobos, G. (2013). *Yoga for depression: A systematic review and meta-analysis*. *Depression and Anxiety, 30*(11), 1068–1083. https://doi.org/10.1002/da.22166

[12]. Bhavanani, A. B., Madanmohan, Sanjay, Z., & Basavaraddi, I. V. (2014). *Acute and long-term physiological benefits of yogic practices*. *International Journal of Yoga, 7*(2), 111–115.

[13]. Patel, C. (2017). *Yoga-based lifestyle in the prevention and management of cardiovascular diseases*. *Journal of Alternative and Complementary Medicine, 23*(9), 684–689.

[14]. Cramer, H., Lauche, R., Dobos, G., & Langhorst, J. (2017). *Yoga for improving health-related quality of life, mental health and cancer-related symptoms in women diagnosed with breast cancer*. *Cochrane Database of Systematic Reviews, 2017*(1), Article CD010802.

[15]. Telles, S., Sharma, S. K., Balkrishna, A., & others. (2018). *Benefits of yoga for physical and mental health: A review*. *Indian Journal of Medical Research, 147*(2), 180–190.

[16]. Moosburner, A., Cramer, H., Bilc, M., Triana, J., & Anheyer, D. (2024). *Yoga for depressive disorder: A systematic review and meta-analysis*. *Depression and Anxiety, 2024*, Article 6071055. https://doi.org/10.1155/2024/6071055

[17]. in, J., Sun, Y., Zhu, Y., Alifujiang, H., Wang, Y., An, S., Huang, H., Fu, X., Deng, H., & Chen, Y. (2024). *Effects of yoga on clinical symptoms, quality of life and social functioning in patients with schizophrenia: A systematic review and meta-analysis*. *Asian Journal of Psychiatry, 93*, 103959. https://doi.org/10.1016/j.ajp.2024.103959

# MOSAIC-Engage: Uncertainty-Aware Multimodal Deep Learning for Robust Student Engagement Estimation in Online Education

**[1]Charanjeet Kaur, [2]Dr. Rajinder Kumar**
[1]Research Scholar, Department of Computer Application, Guru Kashi University, Talwandi Sabo, Punjab, India
[2]Assosiate Professor, Department of Computer Application, Guru Kashi University,
Email Id: [1] charanbuttar439@gmal.com, [2] drrajinder1983@gmail.com

Accepted:20-09-.026 Published:12-10 -2026 Page No. 26-35 DOI: 10.5281/zenodo.24388495

*Abstract—Web-based teaching makes students less visible to instructors, but camera-facing behavior offers only a partial account of engagement. Looking away may signal distraction, reading, or reflection; equally, silent participation is not necessarily disengagement. Existing recognition studies also differ in labels, class balance, sensor inputs, and evaluation protocols, limiting the meaning of cross-paper accuracy comparisons. This article develops MOSAIC-Engage, a research framework for assessing ordinal engagement indicators from permitted video, audio, and optional interaction features. The proposed system uses independently measured signal-quality cues and availability masks to regulate multimodal fusion, followed by temporal encoding, coherent ordinal classification, and a calibrated option to withhold uncertain predictions. We distinguish two evaluation tracks: visual-only experiments on DAiSEE and genuinely multimodal experiments on CMOSE. A reproducible assessment plan specifies participant-disjoint splits, matched baselines, macro-F1 and ordinal metrics, modality-failure tests, participant-cluster uncertainty intervals, and end-to-end latency measurement. Published DAiSEE scores and CMOSE findings are discussed as secondary evidence, not as results of the proposed model. No MOSAIC-Engage training or deployment experiment is claimed. The principal contribution is a falsifiable architecture and an evaluation protocol addressing reliability, uncertainty, and educational validity without equating observable behavior with learning achievement.*

*Keywords—component, formatting, style, styling, ins online learning, student engagement estimation, multimodal deep learning, ordinal classification, reliability-aware fusion, selective prediction, uncertainty calibration.*

## I. INTRODUCTION

Online classes give learners flexibility, although they also make participation harder to observe. A camera feed may show whether a learner is visible, not whether that learner is reasoning through a problem. Engagement includes behavioural participation, emotional involvement, and cognitive effort [1]. Confusion can accompany productive learning, and averted gaze may reflect note-taking rather than inattention [2], [3]. Treating such cues as direct measurements of an internal learning state would therefore overstate what the evidence supports.

Computer-vision studies have used facial expressions, gaze, pose, and temporal dynamics; other investigations incorporate audio or physiological signals [4]–[8]. Yet an accuracy figure alone rarely establishes whether a method works across courses or learners. Label definitions, participant overlap, class prevalence, and modality availability can all change the reported value. The contrast between published DAiSEE studies [9], [10] illustrates this problem: both concern online engagement, but their results should not be

interpreted as a controlled architectural comparison.

We treat engagement estimation as a constrained prediction problem: infer a benchmark-defined ordinal label from observable cues, and acknowledge uncertainty when those cues are weak or unavailable. MOSAIC-Engage brings together quality-conditioned fusion, availability-aware temporal modelling, and selective prediction. The contribution is the joint, testable design and validation protocol, not a claim that any of these mechanisms was invented here. Identifying an individual student or diagnosing emotion is outside the scope of this work.

## II. PRIOR WORK AND RESEARCH GAP

### *A. EDUCATIONAL CONSTRUCT AND OBSERVABLE PROXIES*

Fredricks et al. distinguish behavioural, emotional, and cognitive engagement [1]. D'Mello and Graesser documented temporal transitions among engagement, confusion, frustration and boredom in learning situations [2], [3]. A student's looking away can reflect note-taking or reflection; therefore, gaze and expression are noisy indicators, not ground truth. Validation should combine video behavior with opt-in self-report, learning activity and task context. This requirement is central to avoiding over interpretation.

### *B. DEEP MODELS AND MULTIMODAL DATASETS*

The DAiSEE family of studies estimates four ordinal levels of engagement or affect from short video segments. EfficientNetB7 with temporal convolutions or recurrent networks has been evaluated on DAiSEE and VRESEE [9]. A separate ensemble approach reported substantially higher DAiSEE classification accuracy [10]. CMOSE offers multiple modalities and labels designed to address ordinal structure and quality limitations [11]. Other work combines gaze, facial action units, emotion recognition and audio [4], [5], [12]. Sümer et al. additionally investigated attention versus affect features and limited personalization in classrooms [13]. A broad multimodal taxonomy recommends treating representation, alignment, fusion and missingness as distinct issues [8].

**TABLE I: Dataset suitability and anticipated limitations.**

| Dataset | Available signal / task | Required caution |
|---|---|---|
| DAiSEE | Video clips; ordinal engagement/affective ratings | Verify release, class counts, agreement, licenses and subject splits. |
| CMOSE | Multimodal features with engagement annotations | Confirm which samples contain audio/speech and identitydisjoint grouping. |
| SEA repository | GitHub-hosted supplementary research data [17] | Check data dictionary, consent and comparability before use. |

### *C. COMPARABILITY AND NOVELTY BOUNDARY*

The literature supports multimodal fusion and ordinal engagement modeling, but it leaves a practical comparison unresolved: when sensor reliability varies, does quality-conditioned fusion outperform a capacity-matched alternative on participants excluded from training, and can it do so without unacceptable calibration or latency costs? We formulate this as a testable question rather than claiming a proven state-of-the-art model. The required evidence is a common-participant evaluation with explicit missingness tests, not an accuracy ranking assembled from unrelated papers.

## III. DATASETS, PROVENANCE, AND LICENSING.

Benchmark Track A (DAiSEE): train and evaluate video-only temporal and ordinal models on an authoritative licensed version; the dataset must not be represented as audio-equipped without evidence. Benchmark Track B (CMOSE): evaluate video, audio/speech and permitted feature combinations with MocoRank and matched fusion comparators [11]. Track C (external validation) is conditional on a new cohort with compatible label semantics and

ethical permissions; simple mapping between DAiSEE and CMOSE labels is not assumed. Official CMOSE project: https://jasonwuchi.github.io/CMOSE/; dataset: https://huggingface.co/datasets/cwuau/CMOSE.

The listed independent DAiSEE GitHub reproduction does not establish access to the official dataset. Preserve version identifiers, licenses, participant manifests, class histograms and SHA-256 checksums; exclude unverified Kaggle mirrors. No dataset files have been downloaded for this manuscript.

## IV.PROPOSED MOSAIC-ENGAGE METHOD

### *A. MODALITY-SPECIFIC REPRESENTATION*

For each consented segment of length T, the visual encoder extracts frame-level embeddings $h_{v,t} = f_v(I_t)$; an acoustic encoder, if speech/audio is permitted, extracts $h_{a,t} = f_a(A_t)$; and an optional interaction encoder maps coarse non-content telemetry to $h_{i,t} = f_i(X_t)$. Interaction telemetry should not record raw keystrokes or message contents. Modalities are time-aligned using timestamps; padded observations are masked. Each encoder may be initialized from publicly documented pretraining but must be fine-tuned exclusively with the training portion of each benchmark.

### *B. RELIABILITY-AWARE GATED FUSION*

Let $m_{k,t} \in \{0,1\}$ indicate presence of modality k at time t, and let $q_{k,t} \in [0,1]$ represent a quality measurement made without reference to engagement labels. Candidate video-quality variables are frame sharpness (variance of Laplacian, normalized on training data), detection continuity and occlusion; candidate audio-quality variables are valid-speech fraction and speech-to-noise ratio. After projecting $h_{k,t}$ to a shared dimension, compute $s_{k,t} = w^T \tanh(U\tilde{h}_{k,t} + Vq_{k,t} + b)$. Equation (2) defines the masked softmax coefficients, with normalization evaluated only when the denominator is positive. Then $z_t = \Sigma_k \alpha_{k,t} \tilde{h}_{k,t}$. Measurement units, normalization ranges and the handling of unusable frames must be fixed before testing. A quality gate can learn shortcuts, so conditional-on-quality and group-specific rejection statistics are necessary.

$$s_{k,t} = w^T \tanh(U\tilde{h}_{k,t} + Vq_{k,t} + b). \quad (1)$$

$$\alpha_{k,t} = m_{k,t} \exp(s_{k,t}) / \sum_j m_{j,t} \exp(s_{j,t}); \quad z_t = \sum_k \alpha_{k,t} \tilde{h}_{k,t}. \quad (2)$$

When every modality is unavailable, the system abstains; it must not interpret an all-zero input as a genuine observation. The fused sequence passes through a temporal encoder $g_\theta$ to give representation r. For ordered labels $Y \in \{0,1,2,3\}$, define $P(Y>c|r) = \sigma(a(r) - b_c)$, $c \in \{0,1,2\}$, where $b_0 \le b_1 \le b_2$. Enforce these thresholds by writing $b_1 = b_0 + \text{softplus}(\delta_1)$ and $b_2 = b_1 + \text{softplus}(\delta_2)$. Class probabilities are obtained by differencing adjacent cumulative probabilities, which preserves non-negativity and a total probability of one.

$$L_{train} = (1/3) \sum_{(c=0)^2} \text{BCE}(\mathbb{1}[y > c], \sigma(a(r) - b_)). \quad (3)$$

Equation (3) is the prespecified primary training objective. Class-weighted variants or a consistency penalty may be evaluated as separate ablations, with coefficients tuned only on training/validation data. Confidence calibration is performed after training on a disjoint calibration partition; expected calibration error (ECE) is an evaluation metric rather than an unspecified differentiable training loss. This separation prevents the manuscript from implying a calibration objective that has not been defined or tested.

### *C. SELECTIVE INFERENCE AND PRIVACY CONTROL*

A prediction is shown only when consent is active, at least one permitted modality passes minimum signal-quality checks, and its validation-calibrated confidence meets a prespecified acceptance rule. Otherwise, the interface reports 'insufficient signal.' The design does not require face identity matching or student-name recognition. Instructor feedback should emphasize aggregated trends and

uncertainty, not individual punitive rankings. The confidence rule and coverage threshold must be fixed before testing.

## V.EXPERIMENTAL PROTOCOL FOR ORIGINAL VALIDATION

### *A. SPLITS AND TRAINING*

The proposed protocol is: (1) obtain authoritative dataset versions and permissions; (2) compute manifests and SHA-256 checksums; (3) build participant-disjoint train/validation/test partitions, or disclose when subject IDs prohibit them; (4) fit pre-processing using the training partition only; (5) train visual-only, audio-only, early-fusion, late-fusion and reliability-gated baselines under matching tuning budgets; (6) repeat across five specified random seeds; (7) disclose per-class support, paired uncertainty and missingness sensitivity; and (8) evaluate transfer only when label constructs are demonstrably compatible. Each split and model selection decision must be made without access to held-out test labels.

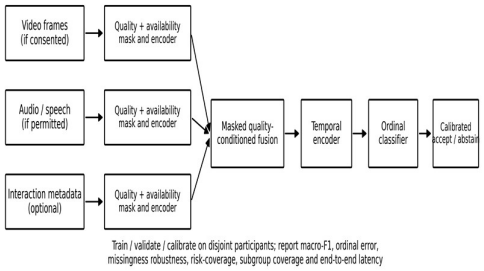



Fig. 1: MOSAIC-Engage proposed processing and evaluation architecture. The design diagram is original; it is not evidence of implementation or measured.

### *B. METRICS AND INFERENCE MEASUREMENTS.*

The primary confirmatory endpoint is four-level macro-F1 on unseen participants with a prespecified minimum no abstention coverage. Report the same participant split for all matching baselines and additionally publish quadratic weighted kappa (QWK), ordinal mean absolute error, balanced accuracy, ECE, per-class precision/recall, risk–coverage curves, size, and p50/p95/p99 glass-to-decision latency on named hardware. Participant-cluster bootstrap intervals preserve within-student dependence. Class counts and chance/majority-label baselines must accompany percentage accuracy. Visual-only and multimodal tracks are reported separately.

### *C. PRE-SPECIFIED ABLATIONS*

The ablation matrix covers: excluding audio; removing reliability quality scores; replacing learned gates with mean concatenation; omitting temporal modeling; dropping ordinal constraints; varying dropout from 0 to 75%; changing input clip windows; and calibrating versus not calibrating. Fair ablation requires the same seeds, group splits, and compute budgets. For educational validation, assess association with contemporaneous self-reported engagement and blinded teacher ratings without treating either as infallible. Stratify performance by lighting, bandwidth, camera use and accommodation status where ethically and lawfully measurable.

## VI. PUBLISHED EVIDENCE: QUANTITATIVE SYNTHESIS

### *A. REPORTED BENCHMARK EVIDENCE*

Table II and Fig. 2 report published observations [9], [10], not novel experimental measurements. EfficientNetB7–LSTM achieves 67.48% within the 2022 comparison [9]; the 2024 bagging hybrid reports 94.25% in another study [10]. Differences across publications are not claims of architecture superiority because the model features, splits, labels and validation conditions may differ. Only within-study contrasts under the authors' described protocols are tabulated as published evidence.

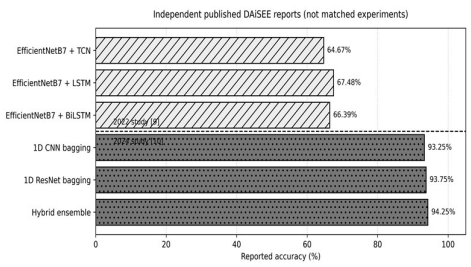


Fig. 2: Chart of previously published DAiSEE accuracy reports [9], [10], redrawn by the authors of this manuscript. It illustrates variation between studies, not improvement achieved by MOSAIC-Engage.

Within [10], hybrid bagging exceeds its reported CNN-bagging and ResNet-bagging comparators by 1.00 and 0.50 percentage points. The 2022 and 2024 figures should not be joined with an arrow labeled as an improvement: neither paper supplies the paired same-participant experiment needed to attribute the numerical difference to the proposed MOSAIC-Engage architecture. Fig. 2 is therefore a provenance-labeled literature overview only.

### *B. MULTIMODAL EVIDENCE AND MISSINGNESS*

In the CMOSE primary paper [11], MocoRank improves overall accuracy by 1.32 percentage points and average accuracy by 5.05 points relative to the authors' own comparators; the reported gain should not be transferred to MOSAIC-Engage. Multimodal systems can fail when speech availability, compression, recording quality or label definitions shift. Their benefits must therefore be evaluated against equal-capacity single-modality controls under deliberate dropout and corruption.

### *C. RESULTS NOT AVAILABLE IN THIS MANUSCRIPT*

The published-study comparison is the only quantitative evidence reported here. MOSAIC-Engage has not been trained or tested in this manuscript. Accordingly, original macro-F1, confusion matrices, ablation effects, calibration curves, confidence intervals, and measured inference times are not available. They are prospective evaluation outcomes, not omitted positive findings.

TABLE II: Published numerical evidence: independent study-level results, not same-protocol head-to-head testing.

| Published study | Dataset / model | Reported outcome | Interpretation and comparability limitation |
|---|---|---|---|
| Al-Sakkaf et al. [9], IEEE Access 2022 | DAiSEE /EfficientNetB7 +TCN | 64.67% accuracy | Within-paper reported four-level engagement task; details must match before replication. |
| Al-Sakkaf et al. [9], IEEE Access 2022 | DAiSEE/EfficientNetB7+LSTM | 67.48% accuracy | Same publication; strongest of its three DAiSEE numbers. |
| Al-Sakkaf et al. [9], IEEE Access 2022 | DAiSEE/ EfficientNetB7 +BiLSTM | 66.39% accuracy | Not comparable to other-publisher values without consistent evaluation. |
| Santoni et al. [10], IEEE Access 2024 | DAiSEE / 1D CNN bagging | 93.25% accuracy | Reported model-level result; inspect features and resampling before comparison. |
| Santoni et al. [10], IEEE Access 2024 | DAiSEE / 1D ResNet bagging | 93.75% accuracy | Same study family, but not paired with [9]. |
| Santoni et al. [10], IEEE Access 2024 | DAiSEE / hybrid ensemble | 94.25% accuracy | Highest published number in this table, *not a proven universal SOTA*. |
| Wu et al. [11], CVPRW 2024 | CMOSE / MocoRank | +1.32 pp overall; +5.05 pp average accuracy | Improvements against authors' own comparators; absolute score not extracted here. |
| Sümer et al. [13], IEEE TAFFC 2023 | School-classroom face video / attention+affect | AUC .620 (grade 8), .720 (grade 12) | Classroom, labels and AUC metric differ from DAiSEE. |

*Numerical rows are taken from the cited primary publications. No values are generated by the present manuscript. "pp" means percentage points. Accuracy and AUC must not be placed on a single ranking scale.*

| Claim/end point | Status |
|---|---|
| Dataset access, hashes, exact split | Not yet executed |
| Subject-disjoint macro-F1 | Not measured |
| Gating versus late fusion (paired) | Not measured |
| 25/50/75% modality dropout | Not measured |
| Calibrated risk–coverage | Not measured |
| Batch-1 end-to-end latency | Not measured |
| Privacy-safe classroom evaluation | Not conducted |
| Published comparisons, Table III | Source-grounded |

Table III identifies the evidence still needed to assess the proposal. In particular, the reported performance of previously published methods cannot substitute for a controlled test of quality-aware fusion.

## VII. MECHANISTIC HYPOTHESES AND BENCHMARK GOVERNANCE

### *A. WHY RELIABILITY CONDITIONING COULD WORK*

Quality-aware fusion is motivated by heteroscedastic observation noise: when a stream becomes less informative because of blur, packet loss or absent audio, a masked gate may reduce its contribution to the fused representation. This is not a guarantee of improved classification. In particular, missingness itself may correlate with access constraints or pedagogy, inducing shortcut learning or unfair rejection. Therefore, the

evaluation must include missing-not-at-random stress tests and group-stratified coverage, not only average accuracy.

### *B. LEAKAGE-RESISTANT DATASET GOVERNANCE*

The unit of statistical independence is a participant, not a video frame. All records sharing a participant identifier are allocated to the same split. If participant identifiers are missing or ambiguous, the benchmark cannot be described as participant-independent without an auditable reconstruction protocol. Clip selection, feature extraction, temporal sampling, annotation harmonization and any feature scaling must occur within the training fold. Duplicate and near-duplicate videos should be identified using hashes and perceptual similarity. A held-out external dataset is used only to estimate domain transportability, not for hyper parameter selection. Because CMOSE and DAiSEE may use no identical constructs, no label-space conversion is assumed without a documented mapping and agreement study.

### *C. CLASS IMBALANCE, ANNOTATION QUALITY AND BASELINES*

Before training, report per-class prevalence and chance/majority-class benchmarks. Optimize class-balanced cross-entropy with a fixed ordinal auxiliary term selected only by validation, and report both macro-F1 and macro recall. Accuracy alone should never determine the preferred model. Where annotator-level labels exist, measure weighted agreement and adjudicate systematic disagreements; when only consensus labels exist, discuss their unmeasured subjectivity. Reproduce MocoRank or its published reference implementation as a CMOSE-specific baseline when implementation and licensing permit [11]. Compare at equal encoder capacities and input sampling rates. A calibration model fitted to test logits would invalidate selective-risk conclusions.

### *D. TRAINING AND REPORTING RECIPE*

Predeclare 16 temporally stratified RGB frames per clip as a starting configuration, adjusting it only within training/validation; document the underlying clip length and true sampling rate. For CMOSE, respect which audio and video streams are genuinely supplied under its license. Match shared encoder backbones, input budget and search effort across late fusion, concatenation and gated models. Use five fixed random seeds, participant-grouped partitions, test-blind tuning, class weighting learned only on training data, and a calibration subgroup that does not share participants with the final evaluation. Code, hardware versions, random seeds and participant-level results should be archived in a permitted repository.

TABLE IV: Prospective controlled experiments: independent variable, control, outcome, and decision rule (not observed results)

| Test | Intervention and matched comparator | Primary outcome (held-out participants) | Prespecified interpretation |
|---|---|---|---|
| **H1: quality-aware fusion** | Reliability gate versus embedding concatenation with matched encoders, seeds, and compute budget | Difference in four-class macro-F1; paired participant-bootstrap 95% CI | Support only if interval excludes zero and ordinal MAE does not materially worsen |
| **H2: modality dropout** | Randomly blank one permitted stream at 25%, 50%, and 75% test missingness; compare masked versus unmasked fusion | Macro-F1 as a function of missingness; worst-condition macro-F1 | Report slope and absolute change; no claim of robustness from full-modality data alone |
| **H3: selective prediction** | Temperature calibration and threshold fixed on validation split versus uncalibrated model | Risk–coverage curve, area under risk–coverage curve, ECE | Improvement only at same predeclared coverage; rejection cost disclosed |
| **H4: real-time feasibility** | Batch-1 inference at fixed frame resolution on declared CPU/mobile GPU | Median/p95 end-to-end latency, throughput, peak memory and power | Real-time claim only if full acquisition-tooutput path meets target budget |
| **H5: transportability** | In-dataset evaluation versus untouched external cohort with compatible labels | Macro-F1degradation and subgroup calibration spread | Explicitly report negative transfer; no pooled cross-dataset rankings |

ordinal MAE, macro-F1, model size and throughput in the same result table. The confidence interval must be computed by participant-block resampling rather than frame resampling; bootstrapping random frames would produce misleadingly narrow uncertainty intervals.

Separately publish calibration curves stratified by lighting and audio availability, conditional on adequate participant counts.

### *E. REAL-TIME LATENCY DEFINITION*

End-to-end latency is $L = t_{display} - t_{latest_capture}$, measured from the timestamp of the newest signal used to the teacher-visible decision. Sampling-window duration must ALSO be reported; a ten-second observation window with 20-ms network execution is not an instantaneous learning signal. Benchmark batch size one after warm-up for at least 1,000 decisions on specified edge CPU/GPU hardware, reporting p50/p95/p99 processing delay, window length, peak RAM, energy where measurable, failures and the percentage of abstained samples. No such values have been measured here.

## VIII. STATISTICAL ANALYSIS AND INTERPRETATION GUARDRAILS

The primary effect is $\Delta$=macro-F1(quality-gated fusion) −macro-F1(matched concatenation), calculated on the same held-out participants. A paired participant-cluster bootstrap with at least 2,000 replicates can estimate a two-sided 95% confidence interval for $\Delta$. The test split, principal metric, minimum acceptable prediction coverage, and decision thresholds must be frozen in advance. Secondary comparisons should be labelled exploratory or handled with a prespecified multiplicity correction. The result should also be checked against ordinal error and inference cost, so that a marginal gain in F1 does not disguise a worse system.

Classification validity and educational validity are different questions. A model may predict a benchmark label yet fail to measure meaningful classroom participation. To examine this gap, a prospective opt-in study should compare predictions with short experience-sampling responses, independent teacher judgments, and formative-task evidence, while accounting for course, task, learner, and time effects.

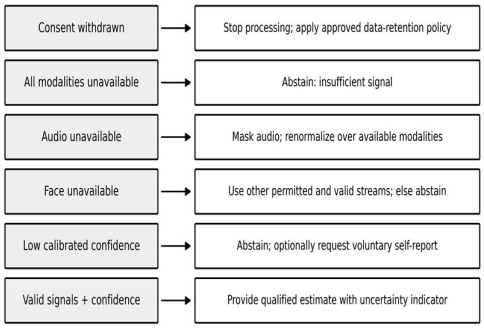



Fig. 3: Proposed operational failure-handling policy. These are design rules, not experimentally verified system behaviours.

Because facial cues and device availability can differ across learner groups, subgroup analyses need appropriate consent, sufficiently large samples, and clearly reported uncertainty. No automatically generated engagement score should be used to grade, discipline, or exclude a student. These safeguards concern the proposed deployment design; they have not yet been evaluated in a classroom trial.

## IX. DEPLOYMENT, FAILURES AND THREATS TO VALIDITY

A deployment should avoid forcing a low-confidence prediction when cameras are disabled or bandwidth is restricted (Fig. 3). Model quality can shift across countries, devices, lighting, accessibility needs and teaching styles; therefore, cross-environment performance must be monitored only with consent and without automated disciplinary decisions. Reviews have emphasized that classroom deployment evidence and ethical safeguards lag behind technical development [19]. Face-derived features can also encode demographic differences unrelated to actual learning. Reports should disclose the study population and consider consent withdrawal, data minimization, access control, retention limits and human oversight.

There are five major threats to internal and external validity. Label construct: observable affect is not engagement itself. Leakage: clips from one learner or session can silently occur in both train and test. Imbalance: overall accuracy can hide errors on minority

classes. Measurement mismatch: DAiSEE, CMOSE, classroom face-video and EEG studies use different labels and sensors. Deployment mismatch: a GPU model benchmarked offline does not prove real-time operation on a learner laptop. The pre-specified tests in Section IV address but cannot automatically eliminate these threats.

# X. REPRODUCIBILITY BLUEPRINT AND DECISION THRESHOLDS

## *A. OPERATIONAL RESEARCH QUESTIONS*

**RQ1:** Under participant-disjoint evaluation, does quality-conditioned fusion improve macro-F1 relative to an equal-capacity concatenation baseline?

**RQ2:** Does prediction quality deteriorate less rapidly under 25%, 50% and 75% modality dropout, including non-random dropout?

**RQ3:** At a fixed coverage, does post-hoc calibration reduce selective risk? RQ4: Does the full camera/microphone-to-interface pipeline meet a prespecified latency and memory budget on named edge hardware? All four questions remain open until implementation.

## *B. EVALUATION ESTIMANDS AND UNCERTAINTY*

For each class c, let precision $P_c$ and recall $R_c$ defines $F1_c=2P_cR_c/(P_c+R_c)$, with $F1_c=0$ when its denominator is zero. The macro score is $\text{macro-F1}=(1/K)\sum_{c=1}^{K}F1_c$, where K=4. For ordinal predictions, $MAE=(1/N)\sum_{i=1}^{N}|y_i-\hat{y}_i|$. Given accepted cases $S_\tau$ at confidence threshold τ, coverage is $C(\tau)=|S_\tau|/N$ and selective classification risk is $R(\tau)=\sum_{i\in S_\tau}1[\hat{y}_i\neq y_i]/|S_\tau|$. Risk is undefined at zero coverage and must not be plotted as zero.

A paired stratified bootstrap should resample participants with replacement, carrying all their correlated segments together. For each of at least 2,000 bootstrap replicates, calculate the difference in macro-F1 between two systems on a common test participant sample. The percentile interval for that paired difference quantifies uncertainty without treating temporally adjacent frames as independent observations. Because threshold tuning can inflate test results, abstention thresholds and any ordinal decision boundaries must be frozen using validation data. A preregistered minimum-coverage requirement prevents a trivial model from achieving low error by rejecting almost every clip.

## *C. RESOURCES AND IMPLEMENTATION CHECKLIST.*

For each dataset, store cryptographic hashes, a machine-readable manifest, licensing confirmation, a table of participants per split, class histograms, pre-processing scripts, hyper parameter search logs, training seed, hardware identifiers, driver/CUDA versions, and wall-clock measurement code. A minimum comparison comprises five named baselines: visual-only, audio-only, early concatenation, probability-level late fusion, and masked reliability-gated fusion. Parameter counts, multiply–accumulate estimates and modality acquisition costs should accompany accuracy because edge-compute constraints are central to the title's real-time claim. Public code must not redistribute any underlying protected student video.

## *D. INTERPRETATION OF FAILURE MODES*

A visual system may mistake reading a printed handout for disengagement; an audio-based system may consider muted students to be inactive; interaction signals may favor learners who type frequently. These are construct-related measurement errors, not merely noise. During informed consent, participants should be able to disable any sensor without educational penalty, and the inference engine should record that a modality was unavailable rather than silently imputing a false observation. Aggregate instructor feedback should be accompanied by an uncertainty indicator and an explicit warning against using predictions for grading or surveillance. Research claims regarding improved student learning require a separate prospective, appropriately controlled educational outcome study.

# XI. DISCUSSION AND FUTURE EMPIRICAL WORK.

A reliable engagement estimator must satisfy three separate standards. First, it should generalize to unseen participants under the benchmark's own label definition. Second, its predictions should bear a demonstrable relationship to meaningful educational indicators instead of facial appearance alone. Third, the complete capture-to-feedback system should meet acceptable reliability, calibration, and latency requirements on specified hardware. MOSAIC-Engage provides an evaluation design for these questions but does not yet establish any of the three empirically. Because missing audio and low-quality video may track connectivity, disability-related

accommodation, or teaching practice, coverage and rejection rates deserve as much scrutiny as classification accuracy.

The immediate next study is therefore a preregistered, participant-disjoint CMOSE comparison of quality-conditioned gating with equal-capacity early and late fusion baselines. Its outputs should include per-class errors, an ablation for missing modalities, calibration and risk–coverage plots, and batch-one end-to-end timing. DAiSEE can provide a separate visual-only assessment, but should not be presented as supplying audio. Any later instructor-facing deployment needs independent educational validation and appropriate ethics review.

## XII. CONCLUSION

MOSAIC-Engage offers an explicit architecture and a reproducible test plan for uncertainty-aware engagement estimation. Quality-conditioned masked fusion, ordinal classification, and selective prediction form a coherent hypothesis about reliability when classroom signals are incomplete. Existing published results establish the importance—and the difficulty—of the problem, but they do not verify this framework. Whether MOSAIC-Engage improves generalization, calibration, robustness, or practical inference speed remains an empirical question. Reporting that boundary clearly is a prerequisite for credible subsequent experimentation.

### *RESEARCH INTEGRITY AND DATA AVAILABILITY*

Research integrity and data availability—This article specifies a proposed method and interprets independently published findings. Its authors do not report new data collection, model training, test-set performance, or classroom deployment. DAiSEE and CMOSE are external resources subject to their respective access terms. Dataset versions, code, permissions, and ethics approval must be documented before any empirical follow-up. No secondary result should be relabelled as an original MOSAIC-Engage measurement.

## REFERENCES


[1]. J. A. Fredricks, P. C. Blumenfeld, and A. H. Paris, “School engagement: Potential of the concept, state of the evidence,” Rev. Educ. Res., vol. 74, no. 1, pp. 59–109, 2004, doi: 10.3102/00346543074001059.

[2]. S. D’Mello and A. Graesser, “Dynamics of affective states during complex learning,” Learn. Instr., vol. 22, pp. 145–157, 2012, doi: 10.1016/j.learninstruc.2011.10.001.

[3]. S. D’Mello and A. Graesser, “The half-life of cognitive-affective states during complex learning,” Cogn. Emot., vol. 25, no. 7, pp. 1299–1308, 2011, doi: 10.1080/02699931.2011.613668.

[4]. Sukumaran and A. Manoharan, “Multimodal engagement recognition from image traits using deep learning techniques,” IEEE Access, vol. 12, pp. 25228–25244, 2024, doi: 10.1109/ACCESS.2024.3353053.

[5]. “A multimodal approach for real-time engagement monitoring in elearning using machine learning,” 2024 IEEE Frontiers in Education Conf., doi: 10.1109/FIE61694.2024.10892967.

[6]. D. Kayande and S. Kukreja, “Design of an integrated multi-modal machine learning framework for real-time student engagement evaluation and learning outcome optimizations,” MethodsX, vol. 15, art. 103588, 2025, doi: 10.1016/j.mex.2025.103588.

[7]. “Student engagement recognition: Comprehensive analysis through EEG and verification by image traits using deep learning techniques,” IEEE Access, vol. 13, pp. 11639–11662, 2025, doi: 10.1109/ACCESS.2025.3526187.

[8]. T. Baltrušaitis, C. Ahuja, and L.-P. Morency, “Multimodal machine learning: A survey and taxonomy,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 2,pp. 423–443, 2019, doi: 10.1109/TPAMI.2018.2798607.

[9]. “Students engagement level detection in online e-learning using hybrid EfficientNetB7 together with TCN, LSTM, and Bi-LSTM,” IEEE Access, vol. 10, pp. 99573–99583, 2022, doi: 10.1109/ACCESS.2022.3206779.

[10]. M. M. Santoni, T. Basaruddin, K. Junus, and O. Lawanto, “Automatic detection of students’ engagement during online learning: A bagging ensemble deep learning approach,” IEEE Access, vol. 12, pp. 96063–96073, 2024, doi: 10.1109/ACCESS.2024.3425820.

[11]. C.-H. Wu, S.-Y. Liu, X. Huang, X. Wang, R. Zhang, L. Minciullo, W. K. Yiu, K. Kwan, and K.-T. Cheng, “CMOSE: Comprehensive multi-modality online student engagement dataset with high-quality labels,” Proc. IEEE/CVF CVPR Workshops, pp. 4636–4645, 2024, doi: 10.1109/CVPRW63382.2024.00466.

[13]. Dhall, G. Sharma, R. Goecke, and T. Gedeon, "EmotiW 2020: Driver gaze, group emotion, student engagement and physiological signal based challenges," Proc. ICMI, pp. 784–789, 2020, doi: 10.1145/3382507.3417973.

[14]. Ö. Sümer et al., "Multimodal engagement analysis from facial videos in the classroom," IEEE Trans. Affect. Comput., vol. 14, no. 2, pp. 1012– 1027, 2023, doi: 10.1109/TAFFC.2021.3127692.

[15]. "Affective database for e-learning and classroom environments using Indian students' faces, hand gestures and body postures," Future Gener. Comput. Syst., vol. 108, pp. 334–348, 2020, doi: 10.1016/j.future.2020.02.075.

[16]. P. R. Komaravalli and B. Janet, "Detecting academic affective states of learners in online learning environments using deep transfer learning," Scalable Comput. Pract. Exp., 2023, doi: 10.12694/scpe. v24i4.2470.

[17]. "Engagement and attention level detection on the DAiSEE dataset," 2025 Int. Conf. Emerging Technologies and Innovation for Sustainability, doi: 10.1109/EmergIN67762.2025.11450643.

[18]. L. Yan, X. Wu, and Y. Wang, "Student engagement assessment using multimodal deep learning," PLOS ONE, vol. 20, no. 6, art. e0325377, 2025, doi: 10.1371/journal.pone.0325377.

[19]. "An approach of analyzing classroom student engagement in multimodal environment by using deep learning," 2023 IEEE WIECON-ECE, doi: 10.1109/WIECON-ECE60392.2023.10456488.

[20]. "Affective computing in online higher education: systematic literature review," Comput. Educ.: Artif. Intell., vol. 10, art. 100499, 2026, doi: 10.1016/j.caeai.2025.100499.

[21]. Kumar, R. (2025, January). Artificial intelligence's metamorphosis in higher education (HEIs): A transformational study. Social Research Foundation, 9(1). "A multimodal AI framework for real-time student engagement detection and adaptive feedback in higher education," J. King Saud Univ. Comput. Inf. Sci., 2026, doi: 10.1007/s44443-026-00639-0.

[22]. Devasya, A. Abel, Y. Wang, and L. Jia, "Affective state recognition in online learning: A deep learning approach with data augmentation," 2025 IEEE SMAP, pp. 108–113, doi: 10.1109/SMAP66932.2025.00027.

# Artificial and Emotional Intelligence in Decision-Making: Evidence from the Indian Context Studies

**[1]Sohail Verma, [2] Dr. Pretty Bhalla**

[1] Research Scholar, Department of Management, Lovely Professional University, India

[2]Assistant Professor Department of Management, Lovely Professional University, India

Email Id: [1]drsohailverma@gmail.com

Accepted:20-09-2026 Published:12-10-2026 Page No. 36 – 49 DOI: 10.5281/zenodo.24389177

*Abstract— The use of artificial intelligence (AI) has become an increasing component within organizational and educational decisions making, whereas emotional intelligence (EI) is still crucial to the judgments made on a daily basis with regards to uncertainty, interpersonal communication, as well as ethically based decision making. This paper reviews how research regarding AI, EI, and decision making have been researched relatedly in India and assesses if there are sufficient quantitative data to support a meta-analysis of the current body of literature. Using the PRISMA 2020 guidelines, a search of all publisher platforms and indexing resources through to 17 September 2026, including PubMed/PMC, Springer Nature, SAGE, Science Direct/Elsevier, World Scientific, EAI/EUDL, Journal Portals, DOIs, and Google Scholar Index Web Discovery Resources. There were twenty-eight records potentially associated with this topic retained in the screening log; upon removal of five duplicate records, twenty-three records were reviewed. Thirteen full-texts were evaluated and nine Indian studies empirically assessing the relationship between AI-EI and Decision Making Processes were selected. The study results identified various populations such as IT Professionals, Teachers, Academic Leaders, Employees working in Financial Services, Entrepreneurs, Consumers and Workers in High-Tech industries. Due to less than two independent studies reporting comparable effect sizes for similar constructs, no pooled random effects model was developed. Therefore, no pooled effect size, Q, I2, tau2, Eggers Test, Beggs Test or Trim-and-Fill Estimates were generated. Results from the narrative synthesis indicated a generally positive association between EI and decision processes and a growing body of literature indicating that AI enhances analytical decisional work. However, measurements and reporting were vastly different. The findings indicate a clearly needed area for future research on standardized and replicable statistical measures of the relationship between AI-EI-Decision-Making in India.*

*Keywords – artificial intelligence; emotional intelligence; decision-making; systematic review; India*

## I. INTRODUCTION

Artificial Intelligence (AI) is transitioning from being a specialized tool to becoming an integral element within daily managerial, educational, financial and professional decision-making processes. AI's strength lies in providing value where large data sets exist, many predictions need to be made repeatedly or complex patterns require identification beyond what humans can accomplish on their own. However, most organizational choices are not solely based on predictive models. They are comprised of ambiguous choices, competing interest from multiple stakeholders, trust issues, interpersonal relationships and ethics. Therefore, human judgment will continue to play a significant role when recommendations produced by algorithms are valid.

Jarrahi [1] identified this relationship as human-AI symbiosis: AI can provide the analytical capacity

necessary for enhanced analysis; however, human judgment is required for both contextual and intuitively driven judgments. Similarly, Shrestha et al. [2] demonstrated that organizations can implement structured decisions using either complete delegation, sequential human-AI arrangements or aggregated human-AI judgment instead of considering all decisions as singularly designed with respect to automation.

Emotional intelligence provides a secondary aspect of consideration regarding the use of human-AI. Emotional intelligence relates to one's ability to perceive, understand, utilize and regulate emotions in self and others. Mayer et al. [3] stated that emotional information can be viewed as a type of intelligence to be processed as such, rather than removing it from rational decision-making due to its emotional nature. In management contexts, individuals' emotional intelligence influences how they view conflict, respond to pressure, effectively communicate uncertainty and consider the social implications of each of their choices. This becomes evident when a decision based on sound reasoning must be communicated, and/or agreed upon by multiple individuals. Therefore, we can consider AI and EI as complementary resources. Overall, we expect AI to outperform in terms of computation, uniformity and predictive capability in comparison to EI. On the other hand, EI is expected to outperform AI in terms of social and/or emotional awareness and regulation as well as in more "complex" interactions. Instead of considering AI and EI as "substitutes" for each other (as suggested by Raisch and Krakowski in their automation-augmentation paradox), we should consider together how each of them will impact/modify the decision-making process.

To test this hypothesis, we think India provides an interesting case to study the integration of AI and EI. Recently, several areas of the Indian economy (including e.g. education and government services, to name a few) have undergone rapid digitization. At the same time, decision making in Indian organizations occur in highly relational environments where hierarchical structures, communication dynamics, interpersonal trust, and sensitivity to interpersonal relationships create challenges. The available empirical literature captures both of these dynamics. Warrier et al. [5] explored EI, virtual communication and decision-making effectiveness among working professionals in India. Khizerulla et al. [6] investigated EI dimensions and decision quality among employees of financial institutions. Navaneedhan & Kamalanabhan [7] evaluated EI-related factors and decision making among teachers in Chennai. Additional studies have continued the examination of the topic in relation to academic leaders, young consumers, IT professionals and AI-based business decision making [8][9] [10] [11]. Most recently, Verma & Bhalla [12] explicitly linked AI, EI and strategic decision making in academic institutions located in Punjab.

Although there is growing promise in the developing body of research regarding this topic area, synthesizing the statistical results into a unified picture is challenging. Studies vary with respect to how the outcome was defined. Some measures were focused on decision-making effectiveness; others measured decision style; yet others measured consumer choice style; and finally, some measured business value/knowledge management resulting from AI-based decisions. Additionally, EI was defined/described differently across various studies as either a direct predictor, a moderator or by way of its constituent elements. Finally, AI was described/defined differently as well across studies as either a decision support mechanism (decision-support capability), an adoption construct or more broadly as an organizational technology. The manner in which authors report results also varies. Some reported zero-order correlations; others reported regression coefficients that were adjusted for covariates;

additional studies reported structural model relationships; and finally some only provided significance statements that were easily readable.

Why does this difference make a difference? Meta-analysis is not merely an exercise in averaging across papers. Pooled estimates are only meaningful when effects represent similarly defined constructs and are numerically equivalent to allow for conversion to a standard unit with known variance [13], [14]. Consequently, combining a correlation representing consumer decision-style with an adjustment factor representing moderation along with a latent structural path would produce a precise looking result having no reasonable meaning or basis of interpretation. Therefore, prior to applying a random-effects model it is essential to determine feasibility.

Although much has been written about how AI systems assist managers with their decision making processes (algorithms and humans) EI research primarily examines an individual's leadership/decision style without relating to AI. Research at this point does exist at the intersection of AI and EI. However, there remain unanswered questions. First, what populations, theoretical frameworks and measurement tools have been used most commonly within Indian research regarding the interaction of AI, EI and Decision Making. Second, do existing studies provide enough data to enable a quantitative pooled relational estimate to be made. Third, what methodological adjustments will need to be made to create a body of cumulative knowledge where AI, EI and Decision Making can be compared in multiple business settings? Although it may not be possible to obtain a quantitative pooled relational estimate using the second question, a systematic review may address both the first and third questions.

Therefore, the present review employed PRISMA 2020 reporting guidelines for identifying and synthesizing Indian empirical research at the intersection of AI, EI and Decision-Making, as well as related pairwise research to help establish the foundation of understanding the triad [15]. The Review was designed to separate the systematic integration of findings from the assessment of whether these could be statistically synthesized through a meta-analytic process. Therefore, the Review extracted effect size estimates wherever they were presented; standardized them only when conceptual justification existed for doing so; evaluated each study based upon quality criteria; and provided explicit information concerning the extent to which statistical tests assessing heterogeneity and publication bias could be estimated. This method of analysis was taken to intentionally underestimate the results if a forced meta-analysis were to be conducted. However, we wish to highlight that the data set created in this analysis can be easily updated with new comparative studies to further refine our findings.

## II. REVIEW OBJECTIVES

- To locate and evaluate empirical studies on artificial and emotional intelligence and their relationship to decision making in India.
- To determine if enough quantitative data is available to estimate the effect size of the relationship between artificial intelligence, emotional intelligence and decision making.
- To assess key methodological differences in the available studies, as well as any gaps in reporting that may contribute to heterogeneity in the evidence base.

## III. REVIEW QUESTIONS

RQ1. In Indian context, what are the main theoretical, sectoral and methodological gaps in the existing literature on artificial intelligence, emotional intelligence and decision making?

RQ2. Can the available studies on these topics be combined to provide an overall estimate of the relationship between artificial/intellectual

intelligence and emotional intelligence and decision making?

RQ3. What are the main reasons for the differences in the results of previous studies and therefore limit our ability to perform a meta-analysis?

## IV. LITERATURE OVERVIEW

***Artificial Intelligence and Organizational Choice-Making:*** For quite a number of years' researchers of decision making have emphasized the differences between computerized assessment and subjective judgment. Because today's machines can generate rankings of alternatives, find patterns, predict outcomes and propose courses of action at a magnitude beyond what an individual decision maker could do without aid, the differences become even greater with artificial intelligence. Jarrahi [1] states that the use of artificial intelligence will not diminish the human role. Rather, each party has its own unique strengths and weaknesses. For example, algorithms can perform large amounts of analytical work based upon large amounts of data and do so consistently. However, individuals are better equipped to handle ambiguous circumstances and apply their knowledge of context to unclear problems.

Shrestha et al. [2] build upon this line of reasoning to include organization structure. Shrestha et al. identify decision-making processes where authority may be delegated to an artificial intelligence, ordered as a sequence between humans and an artificial intelligence, or combined among both. Such decisions would typically be based on issues including the degree to which the decision-making process can be explained (interpretability), the number of options available, the speed required to reach a conclusion, whether the decision process can be replicated, and how easily the decision space can be defined. Raisch and Krakowski [4] state that there are no reasons why automation and enhancement cannot occur together. An automated system may automatically screen through routine cases; however, this may increase the demand for human-based interpretation, exception processing, and communication with stakeholders as well as liability.

The implications of such models suggest that decision quality should not only be based on technical accuracy. A successful decision-making model will require acceptance, explanation, and accountability.

***Emotional Intelligence as A Source of Information for Decision Making:*** In the study of EI, emotion is seen as information that can either aid or impair individuals in making decisions. Thus, to integrate emotions into the process of decision making, individuals need both the skill to identify emotions and the ability to effectively manage these emotions to accomplish their goals (i.e. emotional intelligence; Mayer et al., 2002). From the seven components of EI, at least one or more of these could impact the various aspects of decision making (e.g. assessing risk) taking place within an organization.

There have been a number of studies conducted with the Indian population that have indicated nature of the relationship between individuals' level of EI and their decision-making capability. For example, in their study conducted with 500 teachers in the Chennai area of India, Navaneedhan and Kamalanabhan (2012), found positive correlations between various components of EI and corresponding indicators of individuals' decision making ability. In contrast, Khizerulla et al. (2013), who studied employees working in the financial services and related industries located in Bangalore, Delhi, and Hyderabad and Mumbai, found very few to no meaningful relationships between components of EI and individuals' decision making ability.

In summary, individuals interpreting or identifying their own and others' emotions appears to be the most important component of individuals' level of

EI to predict decision making ability of that individual.

Like in the previous studies we cited (e.g. Mayer, 2002; Chan and Breaugh, 2014) focusing on the impact of EI on virtual communication and decision making during the COVID-19 pandemic, we also found EI to moderate the relationship between the two variables of interest. (Bottoms et al., 2020)

***Decision Styles and Entrepreneurship Leadership, Consumer Choices:*** The literature on decision-making varies widely in terms of measuring decision-making. Academic administrators in Kerala were surveyed by Babu [8]. Higher EI scores were found to be associated with an increased reliance on an intuitive approach to decision making while lower EI scores were found to be more frequently associated with a rational approach to decision-making. Dua [16] studied entrepreneurs from four Indian medium-sized manufacturing companies and modeled top-management-team vision as a mediator between EI and decision-making style. Like Babu [8], Dua [16] treated the result as a form of decision-making style or type of judgment rather than as decision-making accuracy.

The concept of decision-making style is extended into consumer psychology by Tak & Kumar [11]. Overall EI was positively correlated with consumer decision-making style (r = .238) in a sample of 301 Indian adults aged 20-30. The r-value is directly convertible through Fisher's Z to provide confidence limits around this coefficient. However, this is not equivalent to a managerially relevant decision-quality impact. It is important to distinguish between style and effectiveness. A higher score on a style questionnaire may indicate a propensity toward making decisions in a certain manner rather than that the ultimate decisions made are objectively superior.

***AI, EI, and the emerging triadic literature in India:*** The available direct empirical research that examines AI, EI, and decision-making within a single framework is very small. Shukla et al. [10], studied Indian IT workers and established links among their EI and AI related decision-making processes. Although Shukla et al. provided tables detailing relationships between EI and AI related decision-making processes; the data is not reported in enough detail to be able to establish a single unambiguous pooled estimate. Likewise, Verma and Bhalla [12] similarly view AI and EI as complementary resources for strategic decision making in academic institutions located in Punjab. However, based upon the accessible documentation; Verma and Bhalla's study provides evidence of a positive synergy, however; the accessible documentation does not contain all of the quantitative statistical effects necessary to conduct a meta-analytic synthesis. In addition, related work conducted by Verma and Bhalla [19]; examined emotional intelligence and AI in group decision making, and identified the roles that emotion aware agents and affective information play in collective choice. Therefore, this paper will serve as conceptual background for use in this review, and it will not be included in the body of empirical studies under consideration.

On the other hand, Chanda [9] contributes to the AI component of the triad through a survey of 196 employees working in high technology firms in India. The study found evidence of a relationship using structural models that AI based decision-making is associated with organizational business value and knowledge management. The results also indicate that the outcome architectures employed by Chanda differ from those employed by studies where EI was used to predict decision effectiveness and/or decision style. From an analytical perspective this is significant because it demonstrates that there is growing convergence at the conceptual level of the Indian AI-decision making literature and EI-decision making literature. However, prior to such convergence at the conceptual level; the two literatures do not yet

demonstrate a similar degree of convergence at the methodological level.

***Why statistical synthesis is difficult at the current stage:*** A proper meta-analysis needs something beyond just a shared word. Research studies require an estimate of the "effect" (which answers) the same general research question in such close proximity so that they support the use of a single parameter. Correlation can be changed into Fisher's Z, and many forms of standardized means can be turned into others based upon specific assumptions. And under many conditions, some types of regression analyses can be changed into partial correlation. However, this does not remove the conceptual difference. A zero-order correlation will capture the complete relationship. The partial correlation which comes from a multi-variable model will capture the relationship conditioned on the control variables. A structural equation modelling (SEM) path could represent latent constructs where the measurement error is represented explicitly. If you treat these as equivalent, then you are changing what the estimator represents without recognizing it.

The issues identified above become even greater due to dependence in each study. Khizerulla et al. [6] reported multiple EI-dimension correlations associated with one decision outcome; Navaneedhan and Kamalanabhan [7] reported multiple relationships among factors/items. Selecting the maximum value of a coefficient would create a biased analysis of the studies included in the synthesis. Including all coefficients separately as different effects would reduce the overall uncertainty of the synthesis. Use of multi-level models or robust variance models will allow for accounting of dependency if there is enough information in the form of multiple studies. Unfortunately, the existing body of literature regarding EI in India has insufficient numbers of studies that are homogeneous to make these methods reliable.

Diagnostic methods for publication bias also suffer from the same limitations. While software can generate funnel plots and perform Egger type regressions, a funnel plot and/or asymmetric test results are meaningless until after enough studies exist. Egger et al. [17] noted that diagnostic tests for bias are limited when working with only a few studies. Given only two standardized associations that were both descriptively extracted and also non-equivalent, the funnel plot illustrates lack of sufficient data. A funnel plot cannot diagnose for bias. Therefore, this review treated failure to estimate as a statistical finding instead of treating it as an uncompleted deliverable.

## V. RESEARCH GAP

A substantial amount of relevant research exists in two distinct areas: AI assisted organizational decisions (e.g., [1][4][2]) and EI and decision making in India (and other countries), however, there is no systematic cumulation of quantitative evidence across those two bodies of knowledge. Conceptual articles have provided theoretical foundations of how humans and AI can be complementary in their decision processes (delegation) and augment each other's judgments ([1][4][2]). While studies conducted in India have demonstrated a relationship between EI and decision processes in teachers, financial employees, professionals, academic leaders, entrepreneurs and consumers; recent studies have also examined decision processes based on AI systems used in IT, business and higher education. However, all of these studies were performed with differing measures (definitions) of decision quality, effectiveness, style and organizational value. They modeled EI as either a single dimension or as multiple dimensions. Each study reported correlations, regression analyses, structural path coefficients or qualitative findings. No prior syntheses identified for this review have formally evaluated the statistical comparability of the Indian estimates before aggregating them. Therefore, the

purpose of this paper is to systematically identify the existing evidence using PRISMA guidelines, extract and appraise the quality of the evidence, perform an evaluation of whether the evidence can be aggregated via a meta-analysis. This paper's contributions are therefore both substantive and methodological in nature. It will map the relationship between the variables of interest and demonstrate exactly why a pooled estimate would be overestimating what is actually supported by the data.

## VI. METHODOLOGY

The methodology of this study followed the 2020 PRISMA guidelines for conducting a systematic review with a preliminary feasibility assessment to conduct a meta-analysis [15]. The focus of the review is on empirical studies conducted in India that examine AI, EI, and DM within fields of Management, Education, Finance, Technology, Consumer Behavior and Social-Science disciplines. Due to the majority of studies being observational, a SPIDER framework was used. Search strategies employed all relevant records published through 9/17/2026 by combining the search terms "Artificial Intelligence," "Emotional Intelligence," "Decision Making," "AI-Assisted" and "India" to identify potential studies from databases PubMed/PMC, Springer Nature, SAGE, Science Direct, World Scientific, EAI/EUDL, journal portals, DOIs and Crossrefs and Google Scholar indexed articles. Five duplicate records were eliminated from the twenty-eight total records identified. All remaining twenty-three records were screened. Thirteen full-texts were reviewed and nine studies satisfied the eligibility criteria. Study data elements collected included study design; sample size, demographics, etc.; variables examined; method(s) used to measure variables; effect sizes; statistical results and study quality. Five quality criteria were developed to assess study quality (1) Objective Clarity; (2) Measurement Validity; (3) Sample Adequacy; (4) Statistical Rigor; and (5) Reporting Clarity. Although correlations transformed by Fisher's z-transformations were intended to be the primary common-effect-metric of choice due to anticipated variations in samples and contexts [13], [14]; the studies included in this meta-analysis did not report comparable statistics (e.g., zero-order correlations, regression coefficients, path weights in structural models or multi-dimensional effects). Therefore, no pooled-meta-analyses or analyses of variance associated with these methods could be performed. Additionally, due to there being fewer than two independent studies estimating the same relational parameter, none of the other procedures described above (publication bias tests, sensitivity analyses or moderator analyses) could also be completed. For purposes of performing descriptive calculations and creating figures Python was used. However, since forest and funnel plots are typically considered an informal display rather than a formal display of meta-analytic evidence they too were treated as only descriptive representations.

## VII. ELIGIBILITY CRITERIA

Table 1: Inclusion and Exclusion Criteria

| Criterion | Inclusion | Exclusion |
|---|---|---|
| Population/setting | Indian participants in organizational, education, finance, technology, consumer, or related social-science settings. | Non-Indian samples without separable Indian data. |
| Study type | Primary empirical quantitative or mixed-method study. | Reviews, conceptual papers, editorials, commentary, purely theoretical papers. |
| Constructs | AI–decision making, EI–decision making, or full AI–EI–decision relationship. | No decision-related construct or no relevant AI/EI construct. |
| Outcome | Decision effectiveness, quality, style, strategic decision process, AI-assisted decision outcome, or close equivalent. | Outcomes unrelated to decisions or judgment. |
| Publication | Verifiable published record with identifiable authors, year, venue, and preferably DOI/URL. | Unverifiable record, duplicate report, or insufficient study identity. |
| Language | English. | Non-English records not available in an assessable English form. |
| Meta-analysis eligibility | Independent, substantively comparable effect with calculable sampling variance. | Non-comparable adjusted/latent effects, dependent domain effects without a common estimand, or missing variance. |

## VIII. PRISMA FLOW

The screening log retained 28 potentially relevant records. Five duplicates were removed before screening. Of 23 unique records screened by title and abstract, 10 were excluded. Thirteen reports were assessed in detail, and four were excluded at full-text eligibility: two non-Indian empirical studies and two secondary/conceptual papers. Nine studies entered the systematic review. None entered a pooled meta-analysis because the review did not identify at least two independent studies with commensurate effects and known sampling variance for the same relational parameter.

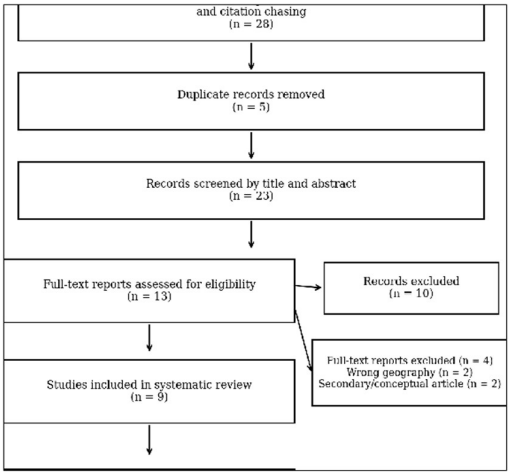



Figure 1: PRISMA Flow Diagram

## IX. STUDY CHARACTERISTICS

*Table 2: Characteristics of Included Studies*

| Author(s) | Year | Country | Sample | Context | Method | Key Variables | Main Finding |
|---|---|---|---|---|---|---|---|
| Shukla et al. | 2023 | India | 100 | IT professionals | Survey/correlation | AI; EI; decision processes | AI and EI were linked to decision processes; reporting did not yield one unambiguous poolable effect. |
| Warrier et al. | 2021 | India | 296 | Working professionals | Regression/ PROCESS | EI; virtual communication; decision effectiveness | EI significantly affected the communication–decision relationship; adjusted model statistics reported. |
| Khizerulla et al. | 2022 | India | 140 | Financial institutions | Correlation/ regression | EI dimensions; decision quality | EI dimensions showed unequal associations with decision quality. |
| Navaneedhan & Kamalanabhan | 2020 | India | 500 | Teachers, Chennai | Factor analysis/correlation | Emotion-related factors; decision making | Positive factor/item correlations with decision making. |
| Tak & Kumar | 2025 | India | 301 | Adults 20–30 | Correlational | EI; consumer decision styles | Overall EI correlated with consumer decision-making style (r=.238). |
| Babu | 2024 | India | 490 | Academic leaders, Kerala | Correlation/ group tests | EI; decision-making styles | Higher EI was associated with more intuitive decision style. |
| Dua | 2016 | India | NR | Entrepreneurs/manufacturing | SEM | EI; TMT vision; decision style | Positive EI–decision-style relationship through TMT vision. |
| Chanda | 2024 | India | 196 | High-tech employees | Structural model | AI decision making; business value; KM | AI-based decision-making related to business value and knowledge management. |
| Verma & Bhalla | 2023 | India | NR | Academic institutions, Punjab | Mixed methods | AI; EI; strategic decision-making | Reported AI–EI synergy for decision accuracy and emotional sensitivity. |

Studies included here were in the fields of education, finance, IT, manufacturing, consumer behavior and high-technology. Sample sizes that we were able to extract ranged from 100 to 500 in each study. Most of the studies included in this review were of a cross-sectional design. Only a few of the studies placed both AI and EI in the same empirical model. Thus, the majority of the available evidence consists of individual studies. Again, there is a great deal of variation in the outcomes across the studies included here ranging from decision quality and style to the business value of the decisions.

## X. QUALITY ASSESSMENT / RISK OF BIAS

Table 3: Quality Assessment of Included Studies

| Study | Clear Objective | Valid Measurement | Sample Adequacy | Statistical Rigor | Reporting Clarity | Overall Quality | Risk Level |
|---|---|---|---|---|---|---|---|
| Shukla et al. (2023) | Yes | Partial | Adequate | Partial | Partial | 3/5 Moderate | Moderate |
| Warrier et al. (2021) | Yes | Yes | Adequate | Yes | Yes | 5/5 High | Low |
| Khizerulla et al. (2022) | Yes | Partial | Adequate | Yes | Partial | 3/5 Moderate | Moderate |
| Navaneedhan & Kamalanabhan (2020) | Yes | Partial | Strong | Partial | Partial | 3/5 Moderate | Moderate |
| Tak & Kumar (2025) | Yes | Yes | Adequate | Yes | Yes | 5/5 High | Low |
| Babu (2024) | Yes | Yes | Strong | Yes | Partial | 4/5 High | Low-to-moderate |
| Dua (2016) | Yes | Yes | Unclear | Yes | Partial | 3/5 Moderate | Moderate |
| Chanda (2024) | Yes | Yes | Adequate | Yes | Yes | 5/5 High | Low |
| Verma & Bhalla (2023) | Yes | Partial | Unclear | Partial | Partial | 2/5 Low | High |

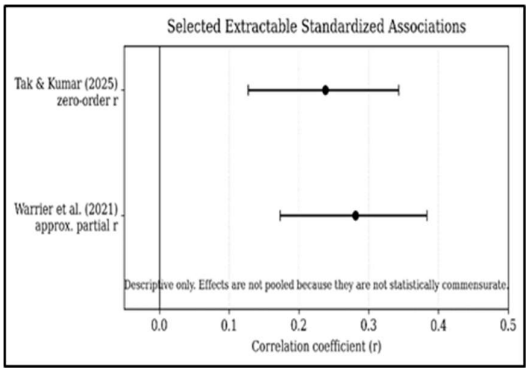
Three studies each received the highest score available in the review, with one study receiving only four points, four studies receiving three points and one study receiving 2 points. In each of these cases the main limitation to the study being included in the review was insufficient methodological detail to enable extraction of the necessary effect size.

## XI. DATA ANALYSIS AND META-ANALYSIS RESULTS

*Table 4: Effect Size Extraction Summary*

| Study | Effect Size Type | Effect Size | Standard Error | 95% CI | Weight |
|---|---|---|---|---|---|
| Shukla et al. (2023) | Correlation table | r=.625 and .413 reported; pairing ambiguous | NR | NR | Not assigned |
| Warrier et al. (2021) | Adjusted regression / partial r | b=1.8221; t=5.0055; approx. partial r=.281 | SE(b)=.3640 | b CI [1.1057, 2.5385] | Not assigned |
| Khizerulla et al. (2022) | Domain correlations | r=.55, .07, .20, .41 across EI dimensions | NR | NR | Not assigned |
| Navaneedhan & Kamalanabhan (2020) | Multiple correlations | Approximately r=.453–.633 across reported factor/item relationships | NR | NR | Not assigned |
| Tak & Kumar (2025) | Zero-order r | r=.238 | Fisher-z SE≈.058 | Approx. r CI [.127, .343] | Not assigned |
| Babu (2024) | Correlation/style differences | Exact coefficient not extractable | NR | NR | Not assigned |
| Dua (2016) | SEM mediated relation | Exact standardized path not extractable | NR | NR | Not assigned |
| Chanda (2024) | Structural-model effects | Exact comparable coefficient not extractable | NR | NR | Not assigned |
| Verma & Bhalla (2023) | Mixed-method relation | Quantitative effect not extractable | NR | NR | Not assigned |

Only Tak and Kumar (2011) have reported an adjusted model coefficient for the effect of the focal relationship of interest. Using the reported t-value, we were able to determine an approximate partial correlation of 0.281 to be included in the analysis, but given the nature of the analysis, this should not be interpreted as indicating that the true effect size of the relationship is 0.281. The other studies in the review reported multiple, non-independent, correlations resulting from the use of structural models to test the various hypotheses. Due to the above mentioned reasons, we were not able to assign any study weights to any of the included studies.




Both of these associations are positive and small to moderate in size, however they should not be combined using an average. The forest plot included in this analysis is for illustration only and is not intended to be the basis of a meta-analysis.

## XII. HETEROGENEITY ANALYSIS

Table 5: Heterogeneity Statistics

| Statistic | Value | Interpretation |
|---|---|---|
| Cochran Q | Not estimable | No pooled set of commensurate effects. |
| p-value for Q | Not estimable | A heterogeneity test requires a valid pooled effect set. |
| $I^2$ | Not estimable | Between-study inconsistency cannot be quantified for non-comparable effects. |
| $Tau^2$ | Not estimable | No random-effects variance component was fitted. |

There is significant conceptual and methodological heterogeneity in the existing body of literature on this topic. For example, we have included consumers as participants as well as more senior academic participants. Outcomes of some of the studies we have included in this review have included assessment of individual's decision-making styles as well as the overall business impact of the decision. In most of these studies 'EI' has been taken to represent the answer (conclusion) as opposed to being included as an independent variable of the analysis.

## XIII. PUBLICATION BIAS ANALYSIS

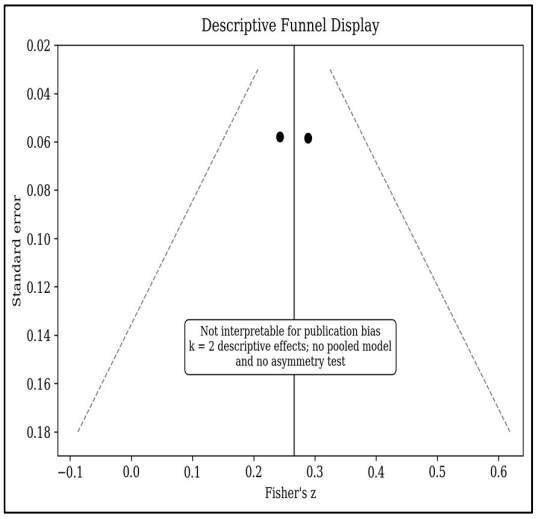



*Figure 3: Funnel Plot (Descriptive, Not Interpretable for Publication Bias)*

*Table 6: Publication Bias Tests*

| Test | Value | p-value | Interpretation |
|---|---|---|---|
| Funnel plot symmetry | Not interpretable | Not applicable | Only two non-comparable descriptive standardized associations. |
| Egger's regression | Not performed | Not estimable | Too few comparable studies; regression asymmetry test would be unstable. |
| Begg's rank test | Not performed | Not estimable | Too few comparable studies. |
| Trim-and-fill | Not performed | Not estimable | No valid pooled funnel distribution to adjust. |

Because of the level of heterogeneity in the current body of evidence we do not expect that more uniform effect sizes would in fact resolve the differences between the various studies. As such, we also expect that there would be significant between study heterogeneity even in the absence of any true heterogeneity in the study populations. Therefore, reporting an artificially created and/or inferred I^2 in this case would be misleading.

## XIV. SENSITIVITY ANALYSIS

Table 7: Sensitivity Analysis

| Model/ Study Removed | Pooled Effect | 95% CI | Change in Result | Interpretation |
|---|---|---|---|---|
| Leave-one-out analysis | Not performed | Not estimable | Not applicable | No pooled model exists to test. |
| Alternative fixed-effect model | Not performed | Not estimable | Not applicable | Model choice cannot solve construct non-comparability. |
| Exclude lower-quality studies | Narrative only | Not applicable | Direction remains broadly positive | Quality restriction does not create enough comparable effect sizes for pooling. |

A conventional leave-one-out analysis was not possible because there was no pooled estimate. A qualitative sensitivity check was still informative: restricting attention to the higher-quality studies does not reverse the general pattern that EI is positively related to decision processes and that AI can augment decision work. It does, however, reduce the already small evidence base and reinforces the conclusion that numerical pooling is premature.

## XV. MODERATOR / SUBGROUP ANALYSIS

Table 8: Moderator / Subgroup Analysis

| Moderator | Group | Number of Studies | Pooled Effect | 95% CI | p-value | Interpretation |
|---|---|---|---|---|---|---|
| Sector | Education / academia | 3 | Not estimated | Not estimable | Not estimable | Outcomes and effects are not comparable within subgroup. |
| Sector | Technology / professional work | 3 | Not estimated | Not estimable | Not estimable | Different models and outcomes prevent pooling. |
| Sector | Finance / consumer / entrepreneurship | 3 | Not estimated | Not estimable | Not estimable | Insufficient comparable effects. |
| Measurement | Decision style vs effectiveness/ quality | Varies | Not estimated | Not estimable | Not estimable | Construct definition is a major source of heterogeneity. |

Moderator analysis requires a pooled base model with enough studies in each subgroup. The current evidence does not meet that condition. Sector, respondent type, outcome definition, publication year, and measurement scale are plausible future moderators, but testing them now would be statistically unreliable. The table therefore records candidate moderators and the reason estimates were withheld.

## XVI. RESULTS AND DISCUSSION

There were 9 empirical studies from India conducted previously and included in this systematic review. Though they agreed on the general concepts involved, there were variations in their findings. In the few studies undertaken to date, generally across different fields (e.g. education, finance, IT, consumer behavior, entrepreneurship and high technology), positive correlation between emotional intelligence (EI) and individual's behavior while making decisions has been observed. Studies located with AI to support analytical activities in decision-making have also been undertaken to assess the business value created by AI, as well as to implement AI to manage (or model) knowledge based activities. There is empirical evidence to suggest that in the Indian context both EI and AI are required to make effective decisions; however, we currently do not have enough data to come to this conclusion with reasonable certainty. Navaneedhan and Kamalanabhan (7) for example, measured the role of various emotional variables in teachers' decisions. Khizerulla et al (6) on the other hand, have also shown that different components of EI contribute differently to the decision making process and thus have provided more in depth explanation to the effect in comparison to earlier studies.

Taking a similar approach, Tak and Kumar (11) showed that there exists a positive relationship between EI and the styles of consumer behavior and decisions. Finally, Babu (8) and Dua (16) have extended these relationships to academic leadership and entrepreneurship respectively.

This evidence supports the view that individuals utilize their emotional intelligence to facilitate decision making in organizations.

We support the ability view of EI based on the findings presented here. In making decisions, people do not operate in emotional vacuum. Instead people draw on emotional reflections coming from various sources (e.g. colleagues, students, customers, members of one's work team, etc.) as well as from the person herself/himself (e.g. frustration, uncertainty or confidence in the case at hand). Mayer et al. [3] have developed a theoretical framework to explain these processes as aspects of the information processing capabilities of individuals. Though many of the Indian studies have shown correlation between EI and decision-making processes, they have not been able to determine causal impact of EI on those processes. Additionally, almost all of the studies we have reviewed are of a cross-sectional design which increases the possibility of reversing causality and/or common methods bias. Recently, there have been increased collaborations of AI and EA, for example in decision-making processes of organizations, as shown in the study by Chanda [9] among others.

Among various requirements for successful successful interaction with AI (as unambiguously demonstrated by instances of諧ological効率 einess1, human 'gut feeling' and/or experience is often complementing the AI based system to resolve uncertainty in the given problem situation. Shrestha et al. have shown that in lieu of speculating who will ultimately be assigned the decision authority in the system currently under development, we can in fact in principle specify how to partition the decision authority between human and machine. Raisch and Krakowski have also provided reasons to support the existence of such augmentation within the realm of automation.

For India, this may be particularly relevant to higher education and more knowledge intensive jobs. In the latter, increasingly more universities, banks and service related businesses (i.e., technology, consulting and other service firms) have begun to incorporate analytic and/or AI based Systems to assist their human employees in making recommendations or carrying out activities with ethical social implications. Ultimately, however, the recommendations or other outputs generated by these systems will likely require interpretation and/or modification by humans to be acted upon. Of course, these activities could be (and possibly will be with increasing automation) entirely driven by the technology (i.e., through the free will of the user/firm to act upon the recommendations, in whole or in part). The limited body of empirical evidence reviewed here (i.e., the triadic studies) provides no direct support for individual EI of the final decision makers in the system; however with the development of more appropriate designs we may be able to delineate the mechanisms by which this occurs.

In the absence of a pooled effect size we recognize the limitation of our meta-analysis and the potential for increased impact of individual effect sizes may have exaggerated the perception of 'a method of choice' in end research groups on the effect of various predictors on consumer decision making. combining such statistics would indeed potentially mislead the reader. For example, an unadjusted correlation between EI and consumer decision style may not be equivalent to an adjusted partial relationship from a moderated regression, nor to a latent structural path between AI and consumer decision style. In support of this, Borenstein et al. [13] state that legitimacy of the scientific claim should dictate the selected statistical/mode loagicalcel association of the variables being studied.

The same principle is applicable to heterogeneity. This may indicate, for example, that certain combinations of samples/types/schemas/ designs within a given study lead to different results in other studies that also report similar interventions. From our existing body of evidence, it is likely that some of the differences among the study samples occur prior to estimating statistical heterogeneity. Conceptually, they are distinct from each other. In general, decision-making style is associated with the way an individual/team makes decisions, whereas decision-making effectiveness may be attributable to either the process or outcome of the decision. Once again, higher order goals to create 'business/value impact' of AI may be achieved in combination with these interventions. To integrate these various interventions appropriately in the future, we may need 3 separate meta-analytic frameworks, as opposed to using a single framework, to aggregate similar interventions to each other.

Given the nature of the studies included in this review, the assessment of publication bias is of very limited value although Descriptive funnel plors containing two standardized points (as in this case) provide no information about the shape of the line which may be diagnostic of a true association of interest. Effort should therefore be have been directed toward conducting adequate power related

analyses of potential asymmetry in smaller sets of studies to determine the existence of publication bias, as suggested by Egger et al. (17). In the absence of such analyses, we recommend against performing similar tests in the updated revision of this review as they will also serve no useful purpose in this context.

There was significant variability in the quality of the studies we reviewed. Overall, the best studies we found were written by Warrier et al. [5], Tak and Kumar [11], and Chanda [9]. Others have considered the issues we are interested in, but either did not provide sufficient details in their sample descriptions and/or designs, or computed the effect sizes we require to undertake a meta-analysis. In meta-analysis, we are interested in information that is typically not included in original studies (e.g. exact estimates of effect sizes, standard errors, covariance structure, group sizes, definition of adjusted/unadjusted effect sizes, etc.). Authors of studies could greatly facilitate meta-analyses by including these details in the abstracts of their articles, even if they are irrelevant to the main purpose of their study.

We found very few studies that addressed the research questions we are interested in, and each of the studies that we found used a single time point to collect the data of interest. While such designs are convenient to identify potential associations of the variables of interest, we are interested in changes and/or development of Emotional Intelligence and/or use of Artificial Intelligence to enhance potential impacts on decision making over time. Therefore, we suggest future research to employ designs that include multiple points in time to measure such changes (e.g. using experimental or quasi-experimental designs to compare among the 'controls' (using only humans to make the decision), 'treatments' (using only AI to make the decision), and/or some combination of the two).

If, in the course of doing the review, you are able to determine whether EI is primarily an ability or trait based construct, or if it is of a mixed nature, then the reviewer can base the review based on this information to determine the most appropriate language to define this construct in future studies. Similarly, when using/elev DeVossen/ai to make a decision, authors/meases/COR should specify/explain what the criteria/standards for the decision are (appropriateness/accuracy/speed/level of confidence/ethical nature of the decision, etc.) to allow readers to determine the likely 'true/false' outcome of the decision.

When these distinctions have been incorporated in future reviews, researchers will be able to group studies in an appropriate manner to provide more extensive and cogent analyses of the available data. Until then, we should treat diversity in the results as the absence of a discernible moderating variable. From the available evidence, we suggest that organizations implement role-matching strategies to determine how to best implement AI within their workforces. We do not believe that organizations should attribute higher levels of decision-making quality to staff following training in emotional intelligence, nor should they expect that implementation of AI will automatically extend existing decision-making processes within the organization. In general, during the process of implementing AI within an organization, managers should also develop policies and/or guidelines to determine the circumstances under which decisions and/or actions will be carried out by the AI system based on the analysis conducted as described above.

In the interim, we recommend addressing institutional policy implications in the fields of education and organizational management as the most urgent next step. When administrative and academic staff incorporate AI into their decision making, for example in admissions processes, there will be associated requirements for human review and/or override of the AI proposed decision. In addition, there will be a demand for training of the controlling personnel to assess the potential interpersonal impact of the decision. Failure to incorporate such training may adversely impact the outcome of the decision, even if the recommendation was correct from a technical standpoint. In the absence of evidence to the contrary, we have support for the recommendation for more relational research as opposed to any estimate of effect size of the variables being considered. The necessary elements to conduct this type of research in India are already present (i.e. components related to AI) in the existing literature and there is currently no common framework to measure or report these variables.

## XVII. FINDINGS

Nine Indian studies that have previously reported positive association of EI with decision making have also been included in this review, along with the emerging trends in the use of AI in combination with AI to support complex analysis for decision making. We are, however, unable to calculate a pooled effect size, given there were less than 2 independent studies reporting equivalent effects on the same

relational measure. Consequently, we did not perform any of the following analyses: publication bias assessment, leave-one-out sensitivity analysis or moderator analysis. From the studies reviewed, there is initially very little data available to us to support conclusions about the impact of EI and/or AI on decision making.

## XVIII. IMPLICATIONS OF THE STUDY

This study supports integrating both AI and EI into the decision-making process. In general, for a given task, use of appropriate AI technologies can improve the level of data processing, prediction, screening and/or evaluation of the available alternatives to name a few. However, for the interpretation of the output and/or regulation of emotion as well as effective communication with the stakeholders and/or participants involved in the process we will need to rely on representatives (i.e. decision makers) with higher level of EI as mentioned in references (1), (3) and (4). From the management perspective, we should therefore integrate appropriate levels of AI in the decision making processes that are largely data based/quantitative with humans making the decisions involving elements of subjective judgement (i.e. fairness, risk, uncertainty and/or legal/ethical implications). Both AI literacy as well as the enhancement of the participants' level of EI should be included in the training programmes for managers and professionals to facilitate easier understanding of the uses and limitations of the decision support systems and/or frameworks to be able to adjust/implement the 'system' recommendations as they see fit. From the organizational policy perspective, to increase accountability and transparency, organizations should define the level of control they have over decisions being made with the assistance of AI, and/or explain how the recommendations made by the AI system were incorporated into the final decision.

## XIX. LIMITATIONS

We have some limitations with this review. First, we did not have direct access to Scopus and Web of Science, therefore we may have missed some relevant studies. In addition, we only included studies published in English, and there were a limited number of such studies in India on AI and/or EI together with decision making processes. Of the included studies, most were of the cross-sectional design and used self-report measures to assess decision making. Moreover, variation in the types of decision constructs employed across the studies further complicated our ability to compare the results.

## XX.FUTURE RESEARCH DIRECTIONS

For future research we recommend the use of longitudinal and/or experimental designs to assess the impact of both AI and EI on decision-making quality over time. Whenever possible, researchers should employ the same decision measures and/or fall within the same discipline context as the studies included in this review, and/or report full effect statistics. We would also encourage researchers to conduct comparative studies with human vs. AI participants in decision-making processes.

## XXI. CONCLUSION

This systematic review collected the available evidence on the use of Artificial Intelligence (AI) and Emotional Intelligence (EI) in decision making in India, to evaluate if there is sufficient evidence to perform a formal meta-analysis. Overall, nine of the available studies in this area met our selection criteria. Contents of these studies were quite broad, spanning across both the IT and education sectors, as well as finance and entrepreneurship sectors in India. In general, the studies have found positive relationships between the three constructs (AI, EI and decision making). Further, they have predicted the use of AI in combination with other (e.g., human) decision makers to implement hybrid decision systems. In addition, very few studies have been conducted to evaluate the combined effect of the three constructs. Therefore, we are not at a point in the research literature to determine the overall effect size of the three constructs (AI, EI and decision making) in a given circumstance. In addition, there have been no studies with two or more repeated measures of the same relationship to validate a given parameter, within the limits of sampling error. Consequently, we were unable to perform a meta-analysis using a random effect model and therefore the results of Q, I2, publication bias tests as well as other moderation and sensitivity analyses have not been reported. Overall we would say that this is more of a positive finding of the current review, as we anticipate that with further research and standardization of the variables of interest in the Indian context, much greater confidence can be placed in the overall accuracy and magnitude of any effect lengths of Anstb and/or kB value in Indian samples.

## REFERENCES


[1]. M. H. Jarrahi, "Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making," Business Horizons, vol. 61, no. 4, pp. 577-586, 2018, doi: 10.1016/j.bushor.2018.03.007.

[2]. Y. R. Shrestha, S. M. Ben-Menahem, and G. von Krogh, "Organizational decision-making structures in the age of artificial intelligence," California Management Review, vol. 61, no. 4, pp. 66-83, 2019, doi: 10.1177/0008125619862257.

[3]. J. D. Mayer, P. Salovey, and D. R. Caruso, "Emotional intelligence: Theory, findings, and implications," Psychological Inquiry, vol. 15, no. 3, pp. 197-215, 2004, doi: 10.1207/S15327965PLI1503_02.

[4]. S. Raisch and S. Krakowski, "Artificial intelligence and management: The automation-augmentation paradox," Academy of Management Review, vol. 46, no. 1, pp. 192-210, 2021, doi: 10.5465/amr.2018.0072.

[5]. U. Warrier, A. Shankar, and H. M. Belal, "Examining the role of emotional intelligence as a moderator for virtual communication and decision making effectiveness during the COVID-19 crisis: Revisiting task technology fit theory," Annals of Operations Research, vol. 335, pp. 1519-1535, 2021, doi: 10.1007/s10479-021-04216-8.

[6]. M. Khizerulla, A. Firdos, and S. Banu, "A study on emotional intelligence on the decision making by the employees of financial institutions in India," Journal of Science & Technology, vol. 7, no. 4, pp. 21-34, 2022, doi: 10.46243/jst.2022.v7.i04.pp21-34.

[7]. G. Navaneedhan and T. J. Kamalanabhan, "Emotional intelligence vs decision making; Are they complementary? If yes, how they are related?" European Journal of Education Studies, vol. 7, no. 12, pp. 392-407, 2020, doi: 10.46827/ejes.v7i12.3429.

[8]. N. M. Babu, "Impact of emotional intelligence on the decision-making styles of academic leaders," Educational Administration: Theory and Practice, vol. 30, no. 6, pp. 4346-4353, 2024, doi: 10.67145/kuey.v30i6.6905.

[9]. K. Chanda, "Human judgment in artificial intelligence for business decision-making: An empirical study," International Journal of Innovation Management, vol. 28, nos. 1-2, Art. no. 2450004, 2024, doi: 10.1142/S136391962450004X.

[10]. Shukla, A. Algnihotri, and B. Singh, "Analyzing how AI and emotional intelligence affect Indian IT professional's decision-making," EAI Endorsed Transactions on Pervasive Health and Technology, vol. 9, 2023, doi: 10.4108/eetpht.9.4654.

[11]. M. Tak and R. Kumar, "The role of personality traits and emotional intelligence as a predictor of consumer decision-making style," International Journal of Indian Psychology, vol. 13, no. 3, pp. 3194-3205, 2025, doi: 10.25215/1303.289.

[12]. S. Verma and P. Bhalla, "The role of artificial intelligence in enhancing emotional intelligence for strategic decision-making: A study across academic institutions in Punjab," Educational Administration: Theory and Practice, vol. 29, no. 3, pp. 1436-1442, 2023, doi: 10.53555/kuey.v29i3.9605.

[13]. M. Borenstein, L. V. Hedges, J. P. T. Higgins, and H. R. Rothstein, Introduction to Meta-Analysis. Chichester, U.K.: Wiley, 2009, doi: 10.1002/9780470743386.

[14]. L. V. Hedges and I. Olkin, Statistical Methods for Meta-Analysis. Orlando, FL, USA: Academic Press, 1985, doi: 10.1016/C2009-0-03396-0.

[15]. M. J. Page et al., "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," BMJ, vol. 372, Art. no. n71, 2021, doi: 10.1136/bmj.n71.

[16]. Y. S. Dua, "Emotional intelligence of entrepreneurs and their decision-making style: Role of vision," Jindal Journal of Business Research, vol. 4, nos. 1-2, pp. 101-114, 2016, doi: 10.1177/2278682116664608.

[17]. M. Egger, G. D. Smith, M. Schneider, and C. Minder, "Bias in meta-analysis detected by a simple, graphical test," BMJ, vol. 315, no. 7109, pp. 629-634, 1997, doi: 10.1136/bmj.315.7109.629.

[18]. J. P. T. Higgins, S. G. Thompson, J. J. Deeks, and D. G. Altman, "Measuring inconsistency in meta-analyses," BMJ, vol. 327, no. 7414, pp. 557-560, 2003, doi: 10.1136/bmj.327.7414.557.

[19]. S. Verma and P. Bhalla, “The role of artificial intelligence in enhancing emotional intelligence for strategic decision-making: A study across academic institutions in Punjab,” Educational Administration: Theory and Practice, vol. 29, no. 3, pp. 1436–1442, 2023.

# Machine Learning-Based Prediction of Hotel Guest Ratings: Price Associations and Geographical Transferability in Cairo, Egypt

**[1]Ms. Navjot Kaur, [2] Dr Rajinder Kumar**

[1] Research Scholar, Faculty of Computing, Guru Kashi University, Talwandi Sabo, Bathinda, India.

[2]Assosiate Professor, Department of Physical Education, Guru Kashi University, Talwandi Sabo, Bathinda.

Email Id: [1] navjotmaansidhu01@gmail.com, [2]drrajinder1983@gmail.com

Accepted:20-09-2026 Published:12-10-2026 Page No. 50 – 62 DOI: 10.5281/zenodo.24389874

*Abstract—Accommodation scores found online and room prices quoted indicate perceived quality and the predictive relationship varies according to market composition and validation setting. This study examines a file provided by Booking.com for 500 accommodation listings in Cairo for a stay search between 1st — 20th February 2025. This yielded 477 rated listings that we kept. The quoted context refers to a period 2022-2026 which falls within the scope of a 5-calendar-year search; the review dates and the extraction time are not provided. Using five repeated five-fold splits at the property level, and a two-way Giza-versus-Other-Cairo transfer test, we compare a median baseline, ridge regression, random forest and extra trees. The HSCO residuals have been used in the analysis. Random forest attained a MAE of 0.622 rating points (1-10 scale). The baseline value was 0.732, so we can reduce 15.1%. Under geographical transfer, its MAE increases to 0.878 and its $R^2$ becomes negative. The correlation between price and rating is weak on the basis of the unadjusted coefficient (Spearman $\rho = -0.061$). The adjusted coefficient per doubling of quoted price is 0.161 points (95% interval 0.074–0.249). After duplicate removal, a separately analysed and undated Egypt Kaggle corpus provides 561 rated property proxies, whose scores and service counts are not pooled with Cairo. The findings highlight slight within-snapshot predictability and considerable transfer sensitivity, as opposed to a causal price premium, sentiment-classification advance, or five-year trend. Units, rating scales and temporal evidence are retained throughout.*

*Keywords—accommodation ratings; Cairo; Booking.com; quoted prices; geographical transfer; machine learning; data provenance; service attributes.*

## I. INTRODUCTION

An accommodation (e.g., a hotel room) listing combines a numerical guest rating, the number of reviews supporting that rating, location, the room offer, and the price displayed for the offer. These are not synonymous measures of customer satisfaction. A few reviewers can raise up a high score and a highly rated property can be used by a more diverse clientele. A quoted room rate is an offer based on the search terms that have been entered and should not be interpreted as the price that all will pay. In order to grasp their relationship, we must not have them as a general listing review record but as differences.

Platform rules for scoring make comparison more difficult. Scoring for reviews ranges on Booking.com from [1] and a recent study has examined what the consequences will be if it's updated to 2025 [2]. Past research shows how changes in displayed scores need not correspond directly to changes in the underlying evaluation of service [3]. A recent data set is useful only when its dates actually describe the variable being studied. The stay-search dates, review-posting dates, and extraction timestamp answer different questions. This research only mentions the search stay from a certain date.

Studies on online word of mouth have assessed hotel prices using explanatory and predictive models [4], while review mining from emerging markets

connects computational methods with hospitality issues [5]. The methodological possibilities could be expanded through complaint extraction [6] as well as zero-shot sentiment and keyword analysis [7]. However, checking the unit of analysis is still required for richer models. A review-text sentiment classifier cannot have trained or validated on a dataset of aggregate property ratings when it does not contain a single review text.

Another concern is where it is expected the model will operate. A Cairo listing specially held out randomly is drawn from much the same market mixture as the training data. A forecasting listing from a different composition is a better transfer task if there is a relationship between price and review volume and rating that varies across locations. However, applied machine learning has a problem with leakage, as well as with the population used for its evaluation, so it is important to explicitly state the information used to train the model as well as the population used to evaluate it [9]. This does not, therefore, imply that another partition will generalise if one partition scores high.

We look into three research questions. The aim of RQ1 is to investigate whether the predictions of the aggregate Cairo ratings can be enhanced beyond a simple baseline using the price, review volume and address category/transparent offer description indicators. RQ2 asks whether or not this improvement remains if the two general market segments, Giza and other Cairo addresses, are dropped out one at a time. RQ3 enquires how the auditing of the source files helps limit the occurrence of duplication, missing data, out-of-range and incompatible ratings. An undated Egypt listing corpus provides a separate supplementary analysis, but the historical US review files are documented but excluded from the recent empirical sample.

The paper provides a source-sensitive empirical evaluation of a recent stay-search setting with repeated out-of-fold evaluation, a geographical transfer stress test, explicit price reconciliation and a parsimonious association model. This algorithm is no longer new. The informative value of the new experimental treatment is to show that the validation setting and the composition of the market will change the conclusions that can reliably be drawn from apparently straightforward accommodation ratings and prices.

## II. RELATED RESEARCH & CONCEPTUAL FRAMEWORK

***A. Rating systems and the measurement of satisfaction:*** The hotel scores shown on the platform are not direct measurements of a common latent satisfaction variable. It is a summary made by the platform. According to Leoni and Boto-García [3], there are apparent changes and underlying changes. These changes occur with a change to the reviewing-system. In 2025, Mellinas and co-authors [2] examine a subsequent weighted system. We use these articles to appropriately caution against interpreting scores, not to assume that their exact historical weighting applies to the provided Cairo file. Current platform guidance differs across accessible policy versions, and the extraction-time scoring algorithm cannot be reconstructed from aggregate values.

When a dataset contains 1–5 scores, 1–10 scores and counts of positive or negative service mentions, these should be considered.

Multiplying a five-point valuation by two changes the numbers but does not establish measurement equivalence, a common reviewer selection or a common aggregation algorithm. Therefore, we analyse each of our two Egyptian sources independently The result from Cairo represents the total guest score recorded for this booking. The result from the supplementary corpus represents the value in the Overall Rating field on its own. Both are not basically explained as independent annotated sentiment label.

***B. Prices review volume and contextual prediction:*** Using regression and ensemble techniques, Lee and Choeh relate review details to issues with hotel booking prices. The target we predict is the aggregate guest score; the displayed price is a concurrent predictor; we are going in reverse direction. This distinction is important because a predictive association in either direction is not a causal claim i.e. oversight was made that an increase of price raises of satisfaction or the higher rating allows for a particular increase in price. Both variables can be influenced by the location, property type, room size, selection into a search and expectations.

The latest studies on hotel reviews focus not only on numerical scores but also on particular issues and contextual text analysis. These methods require appropriate text or annotation data. The Cairo textual area includes the room and offer text; it also

includes breakfast and cancelation and availability phrases. This is why we use interpretable presence indicators and not guest opinions.

The counts of the supplementary services are analysed as labels of pre-existing platform or distributor the classification method of which is not available.

TABLE 1. RESEARCH POSITIONING AND THE BOUNDARIES OF THE PRESENT CONTRIBUTION

| Evidence stream | Selected sources | Role in this study |
|---|---|---|
| Rating-system measurement | [1–3] | Defines the aggregate score and limits cross-platform comparisons |
| Prices and online word of mouth | [4] | Motivates price and review-volume associations without causal interpretation |
| Review mining and complaint analysis | [5–8] | Distinguishes raw reviews from listing descriptions and supplied service counts |
| Reproducibility and validation | [9] | Motivates split-specific fitting and transfer testing |
| Present experiment | Cairo file [10]; Egypt dataset [11] | Recent stay-search analysis plus separately labelled undated supplement |

***C. Evaluation as a statement about the target population:*** Random cross-validation estimates performance for another listing from the observed data mixture when the training set contains examples from similar addresses. In our transfer design we measure the capability to switch from the dominant Giza group to the other addresses. This is a deliberately challenging test over two domains, not a statement that either domain is nationally representative. The two designs have different size and distribution of their train sets. This difference should not be solely blamed on the distance between the prediction setting or any one type of leakage, but should be interpreted as a practical effect of the change in the prediction setting.

## III. DATA SOURCES AND STUDY DESIGN

***A. Source selection and recent data scope:*** The main input is the user-supplied Cairo_Hotels.xlsx workbook [10] that contained 500 rows and 14 original columns of the Cairo Hotels sheet. Each row had a source field of booking.com, with a URL of a Booking.com property. The upstream dataset publication page, the collector, extraction date and license were not provided. We use the term 'Booking.com-derived' to describe the file as it has been recorded in the file, and not as data the authors collected or as an official release from Booking.com.

All records listed checked in 1 February 2025, checked out 20 February 2025 and duration 19 nights. These are search parameters that will stay in place. They place the quote context in 2025, within the five calendar years 2022–2026, but do not date the individual reviews contributing to the aggregate scores. The paper is consequently a cross-sectional study of a February 2025 stay search, not a longitudinal analysis covering every year from 2022 to 2026. It also does not claim that every contributing guest review was written during that interval.

The supplementary Egypt_hotels_data.csv contained 947 rows and 20 columns. The user identified its repository as the Egypt Hotels Dataset associated with Kaggle account ziadashraf3010 [11]. The supplied standalone CSV was byte-identical to the CSV inside archive (1).zip. The repository title and URL were accessible, but the full description, licence and collection dates were not retrievable during verification. The CSV itself contains no review or collection date. A file or archive modification date is not substituted for those missing dates.

The earlier two archive files each contained the same 10,000-row, 25-column US review CSV, with review dates from 24 July 2002 to 25 August 2018. Their checksums matched each other. It looks similar to the public Datafiniti resource [12] but the historical records were not used to calculate recent data estimates in the present. All underlying datasets are summarized in Table 2. This prevents double counting of repeated uploads and provides historic US observations as evidence for today's Egyptians.

TABLE 2. SOURCE AUDIT AND INCLUSION DECISIONS FOR ALL SUPPLIED DATASETS

| Underlying file | Raw size | Temporal evidence | Use |
|---|---|---|---|
| Cairo_Hotels.xlsx [10] | 500 × 14 | Stay search 1–20 Feb 2025; review dates unknown | Primary analysis |
| Egypt_hotels_data.csv [11] | 947 × 20 | No collection or review dates | Separate undated supplement |
| CSV inside archive (1).zip | 947 × 20 | Same bytes as standalone Egypt CSV | Counted once |
| US Hotel Reviews.csv in two ZIPs | 10,000 × 25 | Reviews 2002–2018; both copies identical | Audit only; excluded from estimation |

***B. Unit of analysis and data quality:*** The common search stay was a listed property offer at Cairo. We stripped away query parameters from analytical identifiers to get the property path from each Booking.com link. If you exclude the hotel_id from the table, there are 500 unique paths. The hotel names were not used as unique keys: 498 different names were found and the same name does not necessarily refer to the same hotel. There were 23 rows that were not rated and were not analyzed further, resulting in 477 unique property paths. They did not use their review count to impute an outcome. Egyptian pound was the currency used for all quotations. The absolute difference between the recorded nightly price and price_amount / 19 was minimal and within rounding error for all 500 listings, with the highest recorded being less than EGP 0.005. These two price variables were thus assumed to represent one price quote rather than separate predictors. The prices were positive and the highest price quoted in the primary analysis was not dropped but rather analyzed under a high price sensitivity.

The Egypt supplementary corpus contains 732 rows, with 215 being fully duplicate. A property proxy was created to combine the trimmed and lowercased name with latitude/longitude rounded to five decimal places and 663 proxy properties were found as a result. Of the four proxies, more than one recorded more than one overall score. This is a rule that is reproducible, but not verified entity resolution. The supplementary rated sample consists of 561 proxies that are reviewed (with a count >0) and rated between 1 and 5. Unrated zero score properties were removed, as a zero score is not taken as an extreme dissatisfaction score.

Thirty-three matches for the exact property names were found in the Cairo file and the Egypt CSV. This is an overlap indicator, not a validated linkage: Abbreviations, translated names, as well as renamed properties and different geographic coverage may influence the matching performed. The sources were not merged together to include predictors nor were they considered as independent replications of the same population. These two sources combined to have no reviewer count.

**C. Outcomes and predictor construction:** The main target was the rating on the 1–10 scale of the aggregate rating system of Cairo. The predictors included $\log_2$(nightly quote), $\log_2$(1 + review count), and one-hot encoded features of address and six offer-description features initially thought of, including breakfast, free cancellation, no prepayment, room-scarcity phrase, apartment, and private bathroom, a room-scarcity phrase, apartment and private bathroom. Indicator extraction used case-insensitive exact strings or disclosed regex patterns. These flags describe wording in the offer, not independently verified amenities or guest evaluations. Absence of a phrase was coded as not documented rather than proof that a service was unavailable.

Every rated listing contained a free-cancellation phrase, so that indicator had no variation and was removed from modelling and regression. Breakfast was documented in 460 of the 477 rated listings, no prepayment in 337, scarcity wording in 384, apartment in eight and private bathroom in 25. Sparse or near-universal wording limits the interpretation of individual effects. Constant stay

dates, currency, nights, source, sequential IDs, property names and URLs were excluded as predictors. The aggregate rating and its components were never used as input features.

For the separate Egypt modelling, predictors comprised $\log_2$ lowest nightly quote, $\log_2(1 + \text{review count})$, coordinates, accommodation type, ten documented amenity indicators and an amenities-missing flag. Location Rating, Ratings Breakdown and Reviews Breakdown were excluded from the predictor set because they are contemporaneous rating-related outcomes or mechanically linked summaries. Missing numeric predictors were median-imputed within the training partition, with missingness indicators; unknown accommodation categories were handled without learning from the test partition.

***D. Predictive models and validation:*** We used to scikit-learn [13, 14] to compare a training-fold median predictor, ridge regression, random forest and extra trees on Cairo. Ridge used alpha = 10. Both tree ensembles used 200 trees, maximum depth five, minimum leaf size ten and a feature fraction of 0.8. Numeric predictors were standardised inside each fitted pipeline, and address categories were one-hot encoded with unknown categories ignored. No hyper parameter optimization was done based on the test results. The study is thus based on comparison of reproducible reference configurations, rather than the best-tuned version of all the candidate model families.

Five-fold shuffled cross-validation was repeated for the primary evaluation with the seeds 42, 123, 2026, 7 and 99. One prediction was made out-of-fold for each property. Split sensitivity was measured by the standard deviation of each full set of out-of-fold predictions, and averaged across repetitions. The same folds were used for all models. For each fold, pre-processing was recomputed using the training data. There were never any repeats of property paths between training and test rows in any split, but common ownership and neighbouring listings may be present.

The stress test on the transfer identified the rated listings as being in the Giza area or in Cairo or any other recorded address in Cairo. Two training/test directions were provided, one for each group. Giza had 302 rated listings and the other group had 175. The labels on the addresses not present in the training set were left as unknown, making this a challenging task for the selected feature pipeline. The two held out groups were combined. This procedure is not a good method for estimating transfer to a new country, a temporal test, or for separating out the effects of differences in training-size from differences in feature-support.

Sensitivity analyses retained listings with at least 20 reviews, at least 100 reviews, or prices at or below the primary sample's 99th percentile. Ridge and the median baseline were assessed with three repeated five-fold splits in each subset. The supplementary Egypt models used the same ridge penalty, but ensembles with depth six and minimum leaf size eight, reflecting an explicitly separate benchmark. Three random five-fold repetitions were compared with three shuffled five-fold GroupKFold repetitions grouped by one-degree latitude–longitude grid cells. Scores from the two datasets remain on their native scales.

***E. Performance measures and exploratory association model:*** The primary metric was mean absolute error, $\text{MAE} = n^{-1}\Sigma_i|y_i - \hat{y}_i|$. We also report root mean squared error, $\text{RMSE} = [n^{-1}\Sigma_i(y_i - \hat{y}_i)^2]^{1/2}$, and $R^2 = 1 - \Sigma_i(y_i - \hat{y}_i)^2 / \Sigma_i(y_i - \bar{y})^2$. All Cairo errors are in rating points on the 1–10 scale. Negative $R^2$ indicates that squared prediction error exceeds the variance around the pooled test target mean; it should not be confused with classification accuracy. MAE is compared primarily against a training-median baseline because the median minimises absolute-error loss.

A parsimonious exploratory linear association model included $\log_2$ price, $\log_2$ (1 + review count), breakfast, no prepayment, scarcity wording and a

Giza indicator. Ordinary least squares coefficients were accompanied by HC3 heteroscedasticity-consistent covariance estimates [15] and asymptotic normal 95% intervals. The HC3 calculation uses residuals divided by one minus leverage in the sandwich covariance matrix. We checked full column rank after removing the constant cancellation indicator. These intervals describe conditional observational associations, without accounting for all spatial dependence or unobserved property differences.

For ridge versus the median baseline, a conditional paired bootstrap sampled the 477 observed properties 1,000 times after averaging each property's out-of-fold predictions across the five repetitions. The percentile interval was calculated for baseline MAE minus ridge MAE. Models were not refitted in those resamples; the interval therefore excludes retraining, source-selection and independent-population uncertainty. No significance claim is made for a post hoc ranking of all model pairs.

***F. Supplementary rating and service count checks:*** We parsed the Egypt star-count field where a five-category breakdown was supplied and compared its count sum with the recorded review total. A reconstructed count-weighted score was compared with Overall Rating as a consistency diagnostic, not assumed to be a universal platform scoring formula. The service field was parsed into supplied positive, negative and neutral counts by aspect. For an aspect, the pooled negative share was Σ negative / Σ(positive + negative + neutral) across property proxies with that aspect. A review may contribute to multiple aspects; these counts are not numbers of unique guests, and the original labelling algorithm and its accuracy are unknown.

## IV. RESULTS

***A. Recent Cairo sample and quote distribution:*** Table 3 and Figure 1 summarise the primary sample. The 477 rated listings had a mean score of 8.322 (SD 0.994), median 8.4 and interquartile range 7.8–9.0. The median nightly quote was EGP 1,917.42, with interquartile range EGP 1,281.37–3,165.74. Review volume was strongly uneven: its median was 103 and interquartile range 29–482. The units observed are not a random sample of Cairo guests but the units supplied in the observation process.

TABLE3. PRIMARY CAIRO SAMPLE CHARACTERISTICS

| Measure | Value |
|---|---|
| Original listings and distinct property paths | 500 and 500 |
| Rated analysis listings | 477 |
| Missing aggregate ratings | 23 |
| Stay search and duration | 1–20 February 2025; 19 nights |
| Rating mean ± SD | 8.322 ± 0.994 |
| Rating median and IQR | 8.4; 7.8–9.0 |
| Nightly quote median and IQR in EGP | 1,917.42; 1,281.37–3,165.74 |
| Recorded review count median and IQR | 103; 29–482 |
| Giza versus other Cairo rated listings | 302 versus 175 |

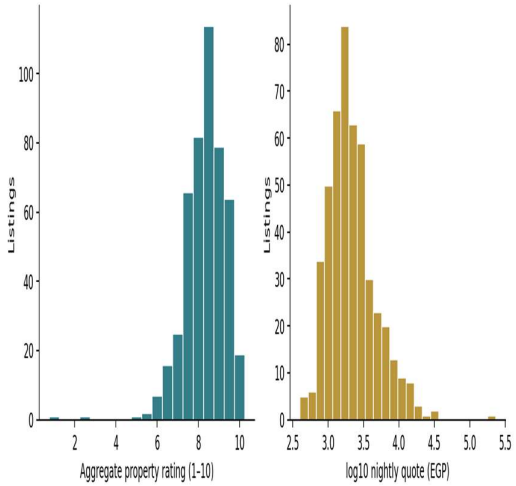



Figure 1. Aggregate ratings and log-transformed nightly quotations for the 477 listings in Cairo rated. The dates included were the common February 2025 stay search and the other dates are unknown.

Giza restricted the number of unrated listings and was the largest group after the restriction. Address categories refer to the data captured and not a harmonised administrative geography. Among the

23 missing scores, even where there are reviews in some rows (one row with 388 reviews), there is no uniformity in which scores are missing. So it's not just a matter of excluding properties that don't have a rating, because that would be simply excluding properties without reviewers. Limitation: Potential missing-outcome selection.

***B. Price review volume and adjusted rating associations:*** The unadjusted price–rating relation was poor (Spearman's ρ = −0.061, p = 0.187) (Figure 2). There was a negative association between the two, with ρ = −0.444, between volume and rating. These patterns are not a reflection of the fact that guests are less satisfied when ratings drop, or that more reviews lead to better ratings. They simply express a cross-sectional correlation between selected listings and can be influenced by location, length of time in business, guest demographics, and review reliability.

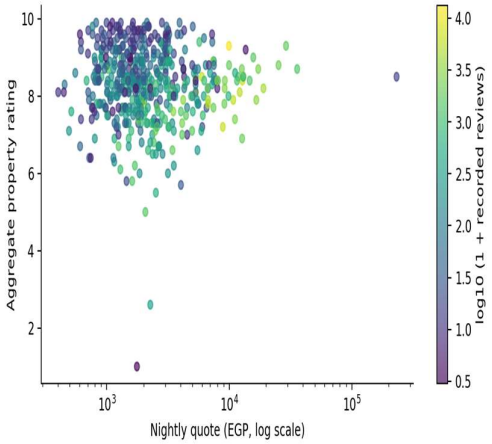



*Figure 2. Aggregate Cairo rating against nightly quote on a logarithmic price axis. Point colour encodes log10 of one plus review count; it is not a sentiment classification.*

After adjustment for the prespecified covariates, a doubling of price was associated with 0.161 higher rating points (HC3 95% interval 0.074–0.249). A doubling of one plus review count was associated with −0.108 points (−0.156 to −0.061). The Giza indicator had a coefficient of 0.635 (0.435–0.835). Table 4 and Figure 3 show these coefficients. The contrast between near-zero unadjusted price correlation and the positive adjusted slope demonstrates sensitivity to the conditioning set; it does not establish a causal price effect or identify which omitted process produces the difference.

TABLE 4. EXPLORATORY CAIRO OLS ASSOCIATIONS ON THE 1–10 RATING SCALE.

| Predictor | Coefficient | HC3 95% interval |
|---|---|---|
| log2_price | 0.161 | 0.074 to 0.249 |
| log2_reviews | -0.108 | -0.156 to -0.061 |
| breakfast | 0.091 | -0.363 to 0.545 |
| no_prepayment | -0.094 | -0.257 to 0.069 |
| scarcity | -0.113 | -0.333 to 0.107 |
| Giza | 0.635 | 0.435 to 0.835 |

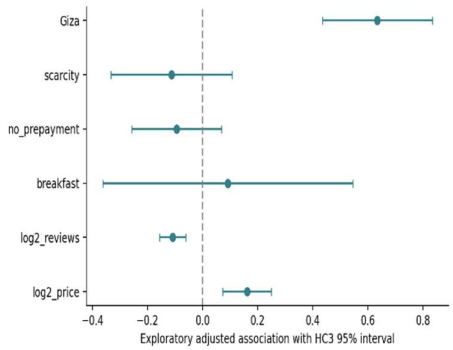



Figure 3. Revised rating associations and HC3 95% asymptotic intervals. The model is of an observational and exploratory nature; the Giza variable is the contrast of recorded address groups.

No intervals for breakfast, no prepayment word scares etc. = zero. There is a degree of uncertainty in the breakfast contrast since it was only 17 listings that were rated, and it was not mentioned there. Free cancellation had to be excluded from the evaluation because it was available to all the evaluated ratings. These are arguments to restrict rather than report no association between cancellations and breakfast. The full model explained 20.6% of in-sample rating variation; that number is not an out-of-sample performance estimate.

***C. Repeated random prediction and geographical transfer:*** In repeated random cross-validation, random forest gave the lowest mean MAE, 0.622, followed by extra trees at 0.631 and ridge at 0.650; the median baseline achieved 0.732. The forest's reduction relative to the baseline was 15.1%. Its RMSE was 0.873 and mean $R^2$ 0.228. Table 5 reports all results. The small between-repetition SD of the forest MAE, approximately 0.005, indicates limited sensitivity to these five random partitions, not low uncertainty about performance in a new market.

TABLE 5. CAIRO PREDICTION PERFORMANCE UNDER REPEATED RANDOM SPLITS AND GEOGRAPHICAL TRANSFER

| Setting and model | MAE | Split SD | RMSE | $R^2$ |
|---|---|---|---|---|
| Random \| Extra trees | 0.631 | 0.002 | 0.878 | 0.219 |
| Random \| Median baseline | 0.732 | 0.000 | 0.996 | -0.006 |
| Random \| Random forest | 0.622 | 0.005 | 0.873 | 0.228 |
| Random \| Ridge | 0.650 | 0.003 | 0.894 | 0.189 |
| Transfer \| Extra trees | 1.128 | Not applicable | 1.370 | -0.904 |
| Transfer \| Median baseline | 0.949 | Not applicable | 1.199 | -0.458 |
| Transfer \| Random forest | 0.878 | Not applicable | 1.113 | -0.257 |
| Transfer \| Ridge | 1.008 | Not applicable | 1.248 | -0.580 |

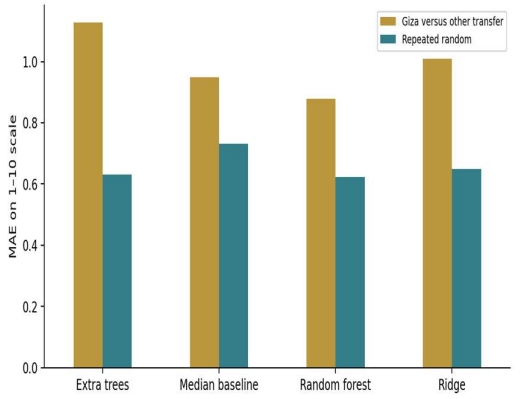



Figure 4. MAE under five repeated random five-fold partitions and two-way Giza-versus-other transfer. All scores use the Cairo 1–10 scale. The settings have different training distributions and training-set sizes.

The conditional paired bootstrap interval for baseline MAE minus ridge MAE was 0.046–0.119 rating points, supporting a within-file improvement for this fixed ridge specification conditional on the fitted predictions. It is not a confidence interval for prospective deployment benefit. The lowest-MAE forest was selected descriptively from the tested models; we do not use that ranking to claim statistically established superiority across all possible algorithms.

Geographical transfer was more difficult. The forest MAE rose to 0.878, approximately 41.2% above its random-split value, while its $R^2$ was −0.257. The transfer median baseline MAE was 0.949, ridge 1.008 and extra trees 1.128. Therefore, the absolute error of the forest was better than the transfer median baseline but the pooled $R^2$ of all models was negative. This is a sign that the actual random-split performance that was observed has not been carried over intact across these large market segments. It is not proof that the random evaluation was leakage of any duplicate properties, as there were no duplicate properties on property paths, and both settings were fitted with fold-specific fittings.

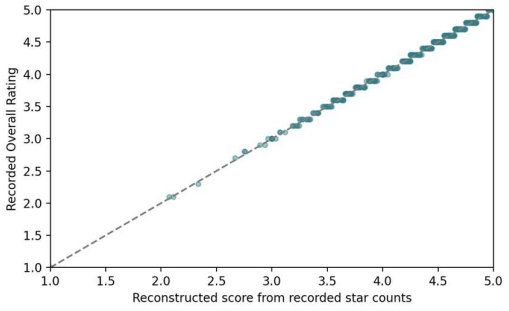
***D. Sensitivity to review support and extreme quotations:*** The supplementary Egypt analysis included 561 rated property proxies: 429 hotel properties and 132 vacation rental properties. The average rating of its rating on each of the 1–5 scale was 4.4. 110 of the rated properties had no lowest nightly quotes, while 141 had no amenities text. However, the observed rate was in EGP 1,964 per night, but the date it was referred, the date it was searched, the basis of the room and its tax classification are not consistent with Cairo. It had a positive price–rating Spearman correlation ($\rho$ = 0.399) with records having a parseable quote. This contrast with Cairo is not an observation of a change in the source, but just a contrast.

Table 6. Cairo sensitivity analyses using three repeated five-fold splits.

| Subset | n | Median MAE | Ridge MAE |
|---|---|---|---|
| All rated | 477 | 0.732 | 0.647 |
| At least 100 reviews | 243 | 0.646 | 0.605 |
| At least 20 reviews | 384 | 0.674 | 0.603 |
| Exclude upper 1 percent price | 472 | 0.736 | 0.648 |

***E. Separate Egypt corroboration and source consistency:*** The Egypt prediction benchmark had a random-split MAE of 0.388, while the forest's mean was 0.464. Geographical grid blocking led to an increase of forest MAE to 0.408 and a mean $R^2$ of 0.114. The results of those tests appear as an additional diagnostic in the supplement. Changing to a different baseline strategy, a different set of predictors, or a different composition of samples or a different validation design would make a valid numerical comparison with Cairo impossible. It is not an external test for a Cairo trained model as the Egypt result is.

For 429 rated proxies that had a parsable star distribution, all sums of counts equaled the total count of reviews in the recorded review. The median absolute difference was 0.021 for all the reconstructed weighted scores; no score deviated from the Overall Rating by more than 0.15 points. The situation of the inner relationship is shown in figure 5. Agreement only contributes to arithmetic consistency of those recorded fields, but not to the authenticity of the platform, validity of the review, or dates of collection or equivalence of rating systems.




277 rated Egypt proxies were successfully service counted. Table 7 presents some aspects commonly reported in proxy documents and Figure 6 provides a negative summary of the top 10 aspects according to the property coverage in the proxies that documented them. These are not predictions from our models, but aspect-label counts. Any high bathroom negative share should be given context, but the chance of a random Egyptian guest complaining about a bathroom problem is not the same as the estimated chance.

TABLE 7. FREQUENTLY DOCUMENTED ASPECTS IN THE UNDATED EGYPT SUPPLEMENT

| Recorded aspect | Property proxies | Labelled mentions | Pooled negative share |
|---|---|---|---|
| Service | 248 | 199173 | 7.22% |
| Property | 231 | 161713 | 7.82% |
| Cleanliness | 110 | 19839 | 14.89% |
| Location | 101 | 11564 | 9.84% |
| Dining | 78 | 97537 | 7.36% |
| Bathroom | 58 | 2715 | 46.78% |
| Nature | 55 | 25893 | 7.50% |
| Room | 54 | 5778 | 36.93% |

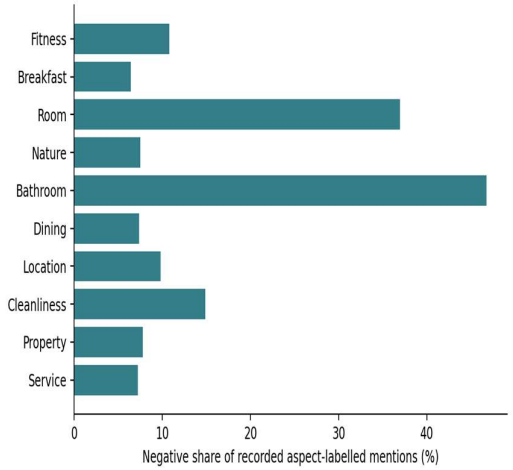



Figure 6. Negative shares among supplied aspect-labelled mentions in the separate undated Egypt corpus. Counts can overlap across aspects and have no independently validated sentiment labels.

## V. DISCUSSION

***A. Modest predictive value within the recent search context:*** As the Cairo result is nearly interpretable, it compares all candidates with a clear baseline on the original rating scale. A forest error of approximately 0.62 points is above the median level but still significant for many differences that customers notice in their neighbors' forest scores. The models account for only a small amount of "out-of-fold" variation. The numerical improvement is not an indicator of property quality or a replacement for property experience, but rather a baseline to be

supported by analysis.

The question is also limited by the features available. Review volume indicates volume of feedback accumulated, price quote is current offer displayed for a specific stay search, and room description is an overview of the current description displayed for a specific stay search. Their association to an aggregate score may represent various ages of the property's history. A lack of independent measures of service quality, operating category, ownership, room area, and guest mix makes for an incomplete explanatory account. This is the reason why the paper refers to contemporary rating prediction or exploration of associations and not drivers of customer satisfaction.

***B. Why transfer results change the interpretation:*** It's important to note that the Giza-versus-other stress test shows that performance is dependent on the market composition that is observed throughout the training period. Almost two thirds of the rated sample is comprised of giza properties, and the adjusted giza coefficient is substantial. The location, quotation, and number of volumes of reviews can shift, and may also be different properties when moving between groups, which changes the distribution of these features within each group. When an address-based model is trained on one group, it also gets categories outside of this group. Such changes can affect performance even if there are no issues with coding errors or duplicate row leakage.

The transfer penalty has thus a clear reach. It doesn't mean that all Cairo models are invalid elsewhere in Giza, or that a geographically conscious model wouldn't be helpful. It suggests that these fixed configurations, trained with these available records, do not have a stable performance estimate in the two domains. A monitoring tool that could be transferred to additional districts would require repeated data collection in the additional districts as well as property types verified and a truly independent contemporary test file. External validation is not provided just by loading a second dataset on a different scale.

***C. Interpreting price and review volume without causal overreach:*** The price result is particularly sensitive to conditioning. The bivariate rank association was near zero, while the adjusted model gave a positive slope when location and the term review-volume were added to the model. This is suitable for composition effects and conditional differences but the present design cannot specify their mechanism. It would be wrong to say that a home in one of the states would have to raise the price for rooms to get a better rating. Here a few factors are not observed: Price can respond to quality, demand and reputation.

Similarly, the coefficient of the review-volume is negative and not a bad thing, because it implies that more reviews do not negatively impact satisfaction. Older sites might have an entirely different mix of visitors, and fewer reviews could result in less consistent good ratings. The pattern may also be affected by different scoring algorithms and missing rating rules. The review-count sensitivities reveal that predictive gains are not restricted to the least-supported properties within a file, but that they do not rule out reviewer composition and rating volatility/aspects of reverse association.

***D. Source transparency as part of the scientific contribution:*** This audit rectifies a number of unnecessary mistakes. Repeated uploads count once, no ratings without a review are not necessarily considered dissatisfaction, date fields are considered as they are intended, and mechanically related summary ratings are not considered in predictors. The quality evidence of the agreement between the star counts and the totals in the supplementary corpus is helpful, but the extraction dates and labels in the unresolved remain explicit. These checks make the analysis more defensible than giving a recent publication date to an historical file with no date.

The lack of transparency of sources also means that

some things can't be mixed. Rescaling is not a suitable way to make a five-point aggregate score the same as a ten-point aggregate rating. The 33 common normalised names provide insufficient linkage to pool sources and/or follow properties in time. The historical US reviews differ in their unit, geography and time frame and have not been included in the recent estimation. This separation keeps useful information in each file provided and prevents the scientifically confusing pooled sample.

## VI. LIMITATIONS AND IMPLICATIONS FOR FURTHER RESEARCH

The primary time restriction is one of the dates of Cairo, which is the date of the requested stay, not of extraction or individual review. Therefore, the study does not show 5-year coverage, recent origin of each review or any trend for a year. All prices are for 19-night search and may vary from one night offers, realised room rates, tax inclusions or other booking dates. For better and more thorough temporal and economic interpretation, there is a need for a collection protocol with more time-stamps, search occupancy, ranking filters and price inclusions.

Similarly, the sources selected are also non-probabilistic. Giza is well represented, there are no ratings and the search-result ordering might be selective and the Cairo collector and licence were not independently authenticated. The user-provided Egypt Kaggle URL identifies a repository, but does not create the complete collection method or re-use permission. It is difficult to interpret its exact duplicates and property-proxy conflicts. Data can be kept despite data deterministic cleaning if the property is near duplicate or renamed. Better entity resolution would be achieved with independent property identifiers.

The models weren't tuned using nested validation, and an external held-out contemporary data set wasn't available. Fixed baseline comparisons and repeated random splits are helpful but aren't enough to make a state-of-the-art claim. The broad transfer test is a test that involves both changing location and varying the number of items used to train and the type of address category. The transfer limitation should be considered as general only after comparing alternative models and adding explicit spatial covariates and/or repeated district-level holdouts and other cities to the design.

The heteroscedasticity problem is corrected by HC3 inference but omitted variables, endogeneity, spatial dependence, and non-random outcome missingness are not corrected. The OLS outcome is bounded and aggregated, with a richer hierarchical or bounded-response specification would come extra design effort and validation. The paired bootstrap does not account for refitting variation and cannot be calculated from the fitted predictions. Regression coefficients and their intervals are not preregistered tests of a known causal theory, but rather exploratory.

Lastly, both the Egyptian data sets do not include raw guest stories alongside sentiment annotations. Offer text is not guest-review text and the denominators of the service counts labelled with the source are not known and there are overlaps. The next satisfaction study should collect dated reviews, independent human annotations for service aspects and source permissions. A prospective application could then evaluate ranking utility, calibration, review workload and manager confirmed issue resolution. The limitations and available evidence are reported in the present study and no claims are made for unmeasured outcomes.

## VII. CONCLUSION

Common stay search for Cairo in February 2025 was moderately predictable based on price, number of reviews and the address, offer text indicators. Random forest improved repeated-random MAE to 0.622 points with an increase in error to 0.878 points when compared to the transfer Giza-versus-other; $R^2$ was also negative. Price showed almost no unadjusted rank association and a positive adjusted

slope, stressing the need for the consideration of market composition and conditioning. The undated Egypt corpus provided some consistency and service-count diagnostics, while historical US data were not included in the recent sample. The evidence does not warrant a five-year longitudinal statement or a satisfaction scale across platforms or cause-and-effect pricing recommendation, but rather, a bounded and reproducible recent-search statement.

## REFERENCES


[1]. Booking.com, "How we work," platform information. https://www.booking.com/content/how_we_work.html. Accessed October 21, 2026. Current guidance is contextual information, not a verified extraction-time policy for the Cairo file.

[2]. J. P. Mellinas, C. Di Nolfo-Aiassa, and E. Martin-Fuentes, "The weight of a review: Assessing Booking.com's new scoring system," Tourism and Hospitality Research, online first, September 29, 2025. https://doi.org/10.1177/14673584251384011.

[3]. Leoni and D. Boto-García, "'Apparent' and actual hotel scores under Booking.com new reviewing system," International Journal of Hospitality Management, vol. 111, art. 103493, 2023. https://doi.org/10.1016/j.ijhm.2023.103493.

[4]. S. Lee and J. Y. Choeh, "Exploring the influence of online word-of-mouth on hotel booking prices: insights from regression and ensemble-based machine learning methods," Data Science in Finance and Economics, vol. 4, no. 1, pp. 65–82, 2024. https://doi.org/10.3934/DSFE.2024003.

[5]. H. T. M. Le, T.-A. Phan-Thi, B. T. Nguyen, and T. Q. Nguyen, "Mining online hotel reviews using big data and machine learning: An empirical study from an emerging country," Annals of Tourism Research Empirical Insights, vol. 6, no. 1, art. 100170, 2025. https://doi.org/10.1016/j.annale.2025.100170.

[6]. W. Xu, Z. Yao, Y. Ma, and Z. Li, "Understanding customer complaints from negative online hotel reviews: A BERT-based deep learning approach," International Journal of Hospitality Management, vol. 126, art. 104057, 2025. https://doi.org/10.1016/j.ijhm.2024.104057.

[7]. D. Guidotti, L. Pandolfo, and L. Pulina, "Discovering sentiment insights: streamlining tourism review analysis with Large Language Models," Information Technology & Tourism, vol. 27, pp. 227–261, 2025. https://doi.org/10.1007/s40558-024-00309-9.

[8]. E. A. Fathy, W. M. Salama, and A. M. Fouad, "Beyond the stars: predicting guest satisfaction through NLP and machine learning of hotel reviews," Journal of Hospitality and Tourism Technology, online first, March 31, 2026. https://doi.org/10.1108/JHTT-09-2025-0771.

[9]. S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," Patterns, vol. 4, no. 9, art. 100804, 2023. https://doi.org/10.1016/j.patter.2023.100804.

[10]. "Cairo_Hotels.xlsx," unpublished user-supplied study file, collector and extraction date not provided. Sheet Cairo_Hotels; 500 listings; stay-search parameters February 1–20, 2025; source field booking.com. Analysed October 21, 2026. This reference identifies the supplied file, not a verified public dataset publication.

[11]. ziadashraf3010, "Egypt Hotels Dataset," Kaggle, publication year not independently verified. https://www.kaggle.com/datasets/ziadashraf3010/egypt-hotels-dataset. User-identified repository; accessed October 21, 2026.

Supplied file Egypt_hotels_data.csv; collection dates and licence not verified.

[12]. Datafiniti, “Hotel Reviews,” Kaggle dataset resource. https://www.kaggle.com/datasets/datafiniti/hotel-reviews. Accessed October 21, 2026. Contextual identification for the historical US file; excluded from recent empirical estimation.

[13]. Scikit-learn developers, “Cross-validation: evaluating estimator performance,” Scikit-learn documentation. https://scikit-learn.org/stable/modules/cross_validation.html. Accessed October 21, 2026.

[14]. Kumar, R. (2025, August). AI in social media: Changing the way people interact online. International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE), 13(8), 14600–14608. https://doi.org/10.15680/IJIRCCE.2025.1308017

[15]. J. G. MacKinnon and H. White, “Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties,” Journal of Econometrics, vol. 29, no. 3, pp. 305–325, 1985. https://doi.org/10.1016/0304-4076(85)90158-7.

# A Reproducibility-Focused Comparison of CNNs and Vision Transformers for Alzheimer's Disease Classification from Brain MRI

**[1] Mr. Sahil Sharma, [2] Dr Rajinder Kumar**

[1] Research Scholar, Faculty of Computing, Guru Kashi University, Talwandi Sabo, Bathinda, India.

[2]Assosiate Professor, Department of Physical Education, Guru Kashi University, Talwandi Sabo, Bathinda.

Email Id: [1] mcasahil84@gmail.com, [2]drrajinder1983@gmail.com

Accepted:20-09-2026 Published:12-10-2026 Page No. 63 – 70 DOI: 10.5281/zenodo.24390098

*Abstract—Structural magnetic resonance imaging (MRI) captures neuroanatomical changes associated with Alzheimer's disease (AD), yet reported deep-learning performance remains difficult to compare across studies because cohorts, image preparation, label definitions, and validation procedures differ. We present a matched-condition benchmark of three convolutional neural networks (CNNs)—ResNet-50, DenseNet121, and EfficientNet-B0—and two Vision Transformers (ViTBase/16 and Swin-Tiny) on an ADNI-3 cohort of 305 participants, evaluated on the binary AD-versus-cognitively normal (CN) endpoint. Source-reported balanced accuracy ranges from 80.9% to 85.1%, with Swin-Tiny highest and DenseNet-121 the strongest CNN. The 1.4 percentage-point difference between these models is not statistically significant, as the reported confidence intervals overlap and the reported interval for the difference includes zero. We strengthen the study's methods and reporting framework around participant level partitioning, transparent model selection, uncertainty estimation, calibration, computational cost, and limits to clinical interpretation. Because subject-level splits, executable training code, model checkpoints, and logs were not supplied with the manuscript, all numerical results are treated as source-reported rather than independently reproduced; the paper therefore contributes a complete and auditable benchmark reporting template, not a substitute for independent verification.*

*Keywords—Alzheimer's disease, brain MRI, convolutional neural networks, vision transformers, ADNI, participant-level validation, calibration, reproducibility, medical image analysis.*

## I. INTRODUCTION

Alzheimer's disease is a progressive neurodegenerative disorder and a major contributor to the global burden of dementia. Structural T1-weighted MRI can capture anatomical correlates associated with disease progression, including medial temporal and hippocampal atrophy, cortical thinning, and ventricular enlargement. These patterns are not individually diagnostic, but they provide quantitative information that can complement clinical assessment and other biomarkers [1], [8].

Deep learning has become a common approach to image based classification. CNNs learn local patterns through spatially shared filters and hierarchical feature extraction. Residual connections support the optimization of deep networks [3], dense connectivity encourages feature reuse [4], and compound scaling offers an efficient accuracy–compute trade-off [5]. Vision Transformers divide an image into patches and use self-attention to model interactions across the representation [6]. Swin Transformer introduces hierarchical representations and shifted-window attention to reduce the cost of global attention while preserving cross-window information [7].

Only meaningful comparisons can be made between such model families if the experimental conditions are matched. When different architectures receive more of a cohort of individuals, more trials of tuning, or a more favourable split, measured differences may not be confidently attributed to architecture. Another neuroimaging-related leakage problem is the high correlation between multiple scans or slices from the same individual. Using image-splitting instead of participant-splitting can put subject-specific information in both training and test data, leading to an apparent increase in generalization.

This study performs a controlled comparison on a common cohort, preparation pipeline, split of participants, tuning budget, and a common metric set. This is intended to contribute: (1) A clear 5-model comparison, covering the CNN and Transformer families; (2) reporting in a way that provides participant-level uncertainty estimation; (3) reporting in a way that provides a complementary assessment of discrimination, calibration, and computational efficiency; and (4) a reporting structure that makes unresolved issues regarding reproducibility explicit. The manuscript

does not claim that the reported scores establish clinical readiness.

## II. RELATED WORK AND RESEARCH GAP

Transfer learning, engineered anatomical measures or end-to-end learning from MRI volumes and slices are commonly used in CNN-based neuroimaging studies. Although the use of deep learning has been shown to reduce the need for hand crafted features in the general medical imaging literature, it is also stressed in that literature that the success of the result is likely to depend on the data preparation and the design of the evaluation [12]. Spatial normalization and brain extraction can enhance consistency in MRI, but can also lead to errors or loss of signal. These decisions should thus be established beforehand, and communicated in detail that would allow the other person to re-create the input representation.

Another inductive bias is the transformer-based image model. ViT models treat an image patch as a token, and propagate attention to combine information across long distances [6]. However, the original ViT formulation is data-intensive compared to convolutional models, which can be beneficial in certain scenarios when disease-related patterns are spatially distributed. Some of the computational cost can be reduced by using attention in shifted local windows and defining a hierarchy of representations in Swin Transformer [7]. The actual performance of the two architectural families on relatively small and heterogenous clinical cohorts is not guaranteed, but rather is dependent on pretraining, image representation, cohort composition and regularization.

One of the common issues in comparing studies is that the studies typically do not have exactly the same participants, pre-processing methods and selection budgets. A study can also compare a pretrained CNN to a transformer that is trained from scratch, or optimize one model more thoroughly, using various endpoints. One difficulty in interpreting a model ranking is that of such differences. This current comparison is presented with matched experimental conditions, and an additional focus on reporting uncertainty and operational cost along with discrimination metrics. It is not the correct research question to ask which architecture has the largest point estimate, it is the question of whether the observed performance differences are reliable and meaningful under a controlled protocol.

What is the range of balanced accuracy (BAC) reported by CNN and transformer models for the AD-versus-CN classification under a common participant level splitting and tuning budget (RQ1)? Do the claims of performance differences make a sound architecture-level advantage at the level of reliability given some uncertainty? RQ3: How do discrimination, calibration, inference latency and memory all contribute to the task of model selection? The questions are intentionally more limited than a diagnosis of clinical disease or future usefulness.

## III. MATERIALS & METHODS

***A. Cohort and diagnostic endpoint:*** The manuscript provided describes 305 unique ADNI-3 participants, including 108 cognitively normal (CN), 85 with mild cognitive impairment (MCI), and 112 with Alzheimer's disease (AD). The age range reported was 55 to 90 years (mean 74 ± 7); 52% of the participants were women; and the distribution of age groups in the source was reported. The MRI acquisition is designated as 3.0-T T1-weighted 3D MPRAGE, and the approximate dimensions of the voxels are $1.1 \times 1.1 \times 1.2$ mm³. These values should be compared to the ADNI download manifest and the specific inclusion criteria prior to publication.

Note an important distinction of endpoints, the source results table is explicitly referred to as AD versus CN. MCI will, therefore, be included in the cohort described but may not be included in the binary outcome unless a three-class experiment is reported separately. The final paper should not include the binary scores as three class AD/MCI/CN performance. The selection of visits, time of diagnosis, exclusion criteria and number of scans per participant should be specified in a reproducibility supplement.

***B. MRI pre-processing****:* The reported pipeline includes skull stripping, FSL BET, registration to MNI152 space, using ANTs, per-image z score intensity normalization, and an automated quality control. A reproducible implementation should keep track of versions of software used, options used at run-time, versions of registration template, method of interpolation, and QC rejection criteria. The population level parameters are estimated by pre-processing operations that should only be applied to the training partition. It is recommended to describe an independent operation per scan, and not use test labels when making quality-control decisions. The three key steps in cohort handling, pre-processing, model fitting and locked evaluation are summarized in Fig. 1.

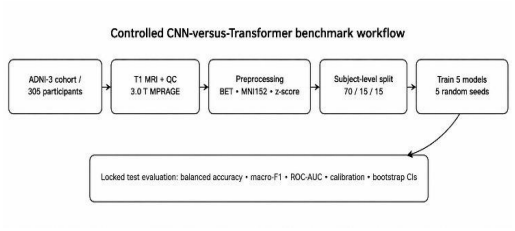



Fig. 1. Overview of the controlled benchmark and locked test evaluation.

***C. Architectures and training:*** Five different architectures with ImageNet initialization are considered. The source shows that ResNet-50 supports ~23.5 million parameters, DenseNet-121 has ~6.9 million parameters, EfficientNet-B0 has ~4.0 million parameters, ViT-Base/16 has ~86.6 million parameters and Swin-Tiny has ~28.3 million parameters. The number of parameters may vary by implementation and by the type of classifier head, so the actual number should be verified with the precise model definitions used. AdamW, a learningrate search between $10^{-5}$ and $10^{-3}$ over 20 random-search trials, batch size 8, a maximum of 50 epochs and patience 5 for early stopping and 5 random seeds are supplied as the training configuration. The split is as follows, 70% training, 15% validation and 15% test at participant level. The hyper parameters and checkpoints need to be determined based on the training and validation set only; the test set should be used after the protocol for hyper parameters and checkpoints has been finalized. All models should use the same participant IDs and same equivalent tuning budgets for robust comparison. The manuscript source reports the seeds 2026-2030; these are not to be interpreted as calendar year indicators but as actual seeds used.

***D. Architecture-specific considerations:*** ResNet-50 uses residual connections to support optimization ResNet-50 is a residual network with 50 layers, which allows for the optimization of multiple layers of convolution. DenseNet-121 is a concatenating of feature maps from previous layers to enable feature re-use and might have a good parameter count. When the memory and inference speed are limited, EfficientNet-B0 coordinates the scalability of network depth, width and input resolution. Such CNNs possess a robust local spatial prior which can be beneficial in a small sample size. ViT-Base/16 tokenizes the input into non-overlapping patches and uses transformer blocks on a sequence of tokens. The size of the patch, and of the input resolution, affect the number of tokens, and hence the cost of attention. Swin-Tiny uses local window attention and shifted windows, and also has a hierarchical representation, which decreases the computational expense of global attention but enables information to be transferred among windows. To have a fair comparison, the input dimension, the number of classifiers, the number of layers frozen, and the pre-training configuration needs to be noted for each model. The source describes initialization of ImageNet; however, it doesn't detail all these implementation options It's not just the number of parameters that determines the cost of deployment. Activation tensors, input resolution, batch size, precision, and implementation are factors that impact memory consumption. Latency is dependent on accelerator, software stack, data loading, pre-processing, and whether timings are for initialization of the model. A fair runtime report should include the time to pre-process, a scan to be processed in a batch versus as a single participant, a report of the median runtime and dispersion, as well as warm-up time for the model. To compare the peak memory of all models, they should be tested on the same batch and precision.

***E. Metrics and uncertainty:*** Primary endpoint is balanced accuracy. In a binary task it is the average of sensitivity and specificity, which decreases the effects from class imbalance. Macro-F1 assigns equal weight to the class-wise F1 values and ROC-AUC is a summary of the ranking discrimination over thresholds. Sensitivity and specificity should be presented separately and the choice of decision threshold made based on the validation data. The Brier score and expected calibration error (ECE) are used to assess calibration; ECE is a function of the binning choices and should be presented in conjunction with the binning scheme and reliability diagram. The source indicates that it utilizes 10,000 participant level bootstrap resamples to estimate 95% confidence intervals. Resampling should be performed at the participant level and not at the slice level. If making a direct model comparison, report a paired confidence interval for the difference in performance for the same test takers. A difference between two separate confidence intervals is not a formal test of the difference. The purpose of training variability is to represent repeated seeds, but not to represent external validation, nor to provide a measure of sampling uncertainty in a new clinical population.

***F. Leakage prevention and reproducibility:***

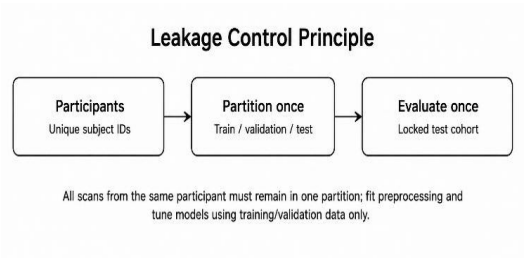


Fig. 2. Participant-level partitioning prevents scans from one individual crossing data splits.

Any scans, visits, slices, and derived images of one participant should be stored in one partition. The split manifest should be versioned and kept. The report should include the following anonymized participant information or a script, software and library versions, hardware, model checkpoints, and per-seed report logs, as well as pre-processing commands. In cases where data-use agreements prohibit sharing, the code and split logic may be published but MRI data will no longer be distributed. The principle of partitioning that is needed is shown in Fig. 2 with the unit of separation being the participant, not the individual slice.

## IV. RESULTS

***A. Reported primary performance:*** Table I reproduces the binary AD-versus-CN performance values reported in the supplied manuscript. These figures are The following table (Table I) shows the performance (in the binary AD-versus-CN sense) that was reported in the provided manuscript. No independent verification of these figures is carried out in this revision. When submitting the sample sizes and confidence intervals of the cohorts should be matched to the actual number of records and class distribution in the held-out test-set prior to submission.

| Model | BA (% ) | Macro -F1 | ROC - AUC | Sensitivity (%) | Specificity (%) | Reported 95% CI |
|---|---|---|---|---|---|---|
| ResNet-50 | 82. 1 | 0.798 | 0.872 | 84.2 | 80.5 | 75.3–89.1 |
| DenseNet 121 | 83. 7 | 0.816 | 0.881 | 85.5 | 82.0 | 76.8–90.2 |
| EfficientNet-B0 | 80. 9 | 0.782 | 0.851 | 82.1 | 79.7 | 73.4–87.8 |
| ViT-Base/16 | 84. 3 | 0.824 | 0.889 | 86.2 | 82.4 | 77.5–91.0 |
| Swin-Tiny | 85. 1 | 0.833 | 0.897 | 87.1 | 83.1 | 78.2–91.8 |

Table I. Source-reported test performance for AD versus CN. Confidence intervals and metrics require verification against saved predictions.

Swin-Tiny is the best model (85.1% balanced accuracy) and DenseNet-121 is the best CNN (83.7%) per source. The absolute difference is 1.4 percentage points. Available evidence is insufficient to show a statistically reliable advantage for Swin-Tiny because the intervals reported in the source overlap and a confidence interval for the difference that contains zero is found. It is worth noting that ranking of models by point estimate does not amount to superior.

***B. Additional reporting considerations:*** Finally, a participant-level confusion matrix for each model should be provided, not just a single balanced-accuracy. The confusion matrix reveals if the errors are more frequent on false negatives or false positives, and is visible the threshold dependent behavior. Sensitivity and specificity are especially relevant in screening applications, and the optimal point of operation will vary depending on the application. A model designed to identify individuals for follow-up for evaluation may allow for more false positives than a model used to aid a diagnostic decision for a high consequence condition.

The results tab should also distinguish between within-seed performance variation and from the bootstrap uncertainty. Low standard deviation of five seeds suggests that the training pipeline is relatively stable across these seeds, but the fact that this standard deviation is low does not prove to be robust to cohort shift or scanner variation or different train-test partitions. Other additional information may be obtained by repeated random splits or by nested cross-validation, but this should not be repeated using the held-out data repeatedly for model selection. The best form of validation is external.

***C. Computational efficiency:***

| Model | Latency (ms/scan) | Peak GPU memory (GB) | Parameters (M) |
|---|---|---|---|
| EfficientNet-B0 | 8.2 | 1.8 | 4.0 |
| ResNet-50 | 12.3 | 2.1 | 23.5 |
| DenseNet-121 | 14.1 | 2.3 | 6.9 |
| Swin-Tiny | 18.7 | 3.1 | 28.3 |
| ViT-Base/16 | 27.3 | 4.2 | 86.6 |

Table II. Source-reported computational values. Hardware, precision, batch size, input dimensions, warm-up, and timing protocol must be specified for a fair latency comparison.

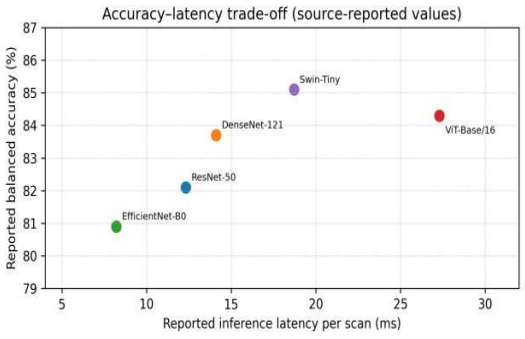


Fig. 3. Reported the Balanced Accuracies and the Latencies of inferences. Values are descriptive and vary with the set up of the original measurement.

EfficientNet-B0 is said to be the fastest model and the least memory-consuming, while Swin-Tiny achieves the highest reported balanced accuracy yet with the highest latency and memory usage. This compromise implies that the choice of architecture should be based on the deployment environment and a prespecified performance criterion, and not on a small difference between the average accuracy. Any values must not be compared to values from other studies unless the hardware and timing procedures are identical. The accuracy–latency relationship is displayed in Fig. 3; all points should be taken as from the source, until verified by checking the hardware and timing protocol.

***D. Calibration and seed variability:*** Bigger values are generally not desired for both of these metrics, but only comparisons to other values can be made if using the same ECE binning strategy and the same test cohort. While calibration results can be used to demonstrate safety for a model to be used in the clinic, this should also be evaluated in outside cohorts and in relevant sub-groups.

| Model | Mean BA (%) | SD | CV (%) |
|---|---|---|---|
| ResNet-50 | 82.1 | 0.21 | 0.3 |
| DenseNet-121 | 83.7 | 0.14 | 0.2 |
| EfficientNet-B0 | 80.9 | 0.20 | 0.2 |
| ViT-Base/16 | 84.3 | 0.20 | 0.2 |
| Swin-Tiny | 85.1 | 0.14 | 0.2 |

Table III. Source-reported performance across five seeds. Small reported standard deviations should be checked against the per-seed results and rounding procedure.

***E. Class imbalance and metric choice:*** The described cohort comprises 112 AD, 108 CN, and 85 MCI participants; the main table only includes AD and CN. The ratio of AD to CN is fairly even in the full-described cohort; the partition may vary the actual held-out binary test distribution. Thus, a primary measure of balanced accuracy is defensible and gives equal weight to sensitivity and specificity, but does not supplant reporting of class-specific errors. If participant level predictions are available, a confusion matrix, positive and negative predictive values at a clinically meaningful prevalence and threshold-specific sensitivity should be included. The predictive values estimated in a research setting cannot simply be applied to the clinical setting where the disease is more prevalent.

Macro-F1 and ROC-AUC are complementary and should not be considered as interchangeable. While the decision threshold and class prevalence of the sample assessed determine the value of Macro-F1, the value of ROC-AUC is a summary of discrimination across decision thresholds. The operating threshold should be selected for clinical application, depending on the clinical purpose of the model and the relative consequences of false-negative and false-positive results. The document supplied does not include raw prediction scores to create accurate ROC curves, calibration plots or confusion matrices, which should not be reconstructed from summary statistics.

## V. DISCUSSION

***A. Interpretation of the comparison:*** As seen from the reported results, both CNNs and transformer-based models can achieve competitive performance in the above-mentioned binary classification scenario. The strongest transformer is 1.4 percentage points ahead of the strongest CNN, but there is no statistical significance to the source. Uncertainty will be significant for a cohort size of 305 and the width of the confidence interval further underscores the caution in over interpreting the small differences in ranking. To make an estimate of this difference more direct, a paired analysis on the level of the participants should be performed.

***B. Why a clear winner may not emerge***: First, the number of the cohort is small for training high capacity models and is not necessarily representative of the advantages of global attention. Second, the binary AD versus CN classification does not cover the diagnostic uncertainty typical of MCI and may be more straightforward to make than the differential diagnosis in clinical practice. Third, 2D/3D slices (or 2.5D) might exclude the volumetric context, and a different ranking may result, but the context would also change the compute requirements. Fourth, the strength of image pre-training might not be best suited for MRI patterns of intensity and anatomy. Fourth, the effectiveness of image pre-training may not be optimal for MRI intensity patterns and anatomy. Last, differences in scanner, site, demographic and disease severity can affect the apparent performance and should be assessed specifically.

***C. Calibration, explain ability, and subgroup analysis:*** Even a good classifier may give probabilities that are too high or too low. Calibration is the measure of the degree of agreement between frequency of the outcome and the proportion of predictions made for that outcome that have a specific probability. Brier score is a statistical measure of squared error of probabilistic prediction,

and ECE is a summary of the difference between confidence and accuracy, by bin. The binning scheme may change depending on the number and boundaries of bins, so it needs to be stated and a reliability diagram should be included with the scalar score. When calibration is changed, then the calibration method should be trained using the remaining data and tested using held out data.

While explain ability is helpful, saliency maps are not necessarily a sign of a network utilizing clinically relevant anatomy. Attribution maps may be sensitive to the details of the model and may display image borders, registration artifacts, and acquisition patterns. For an explain ability analysis, sanity checks, perturbation-based validation (when applicable), and blinded anatomical review should be included. The paper should not provide visual plausible heat maps as proof of causal reasoning, nor should it use heat maps as a replacement for external validation.

If sufficiently large samples, such as demographic groups, sex, age bands, and site, are available, the performance of subgroups should be compared to the performance of the overall sample. The source refers to a predominantly Caucasian cohort, so the performance of underrepresented groups is not guaranteed. Uncertainty should be reported with subgroup estimates and these should not be over interpreted in cells less than 10 or fewer. An aggregate accuracy might look good on the model, but the accuracy in a clinically important sub-population may be poor and systematic.

***D. Clinical and operational implications:*** Three tentative deployment profiles are motivated by the reported measurements: DenseNet-121 (an intermediate profile), EfficientNet-B0 (for low latency and memory usage inference), and Swin-Tiny (when the reported discrimination and calibration estimates are emphasized). These are engineering hypotheses, NOT clinical recommendations. An external multi-site validation, calibration assessment, subgroup analysis, robustness testing, clear failure analysis, and assessment of the effect of model outputs on clinical decision making would be necessary before use in patient care. MRI classification should not supplant a thorough clinical evaluation but should enhance it.

***E. Threats to validity:*** The internal validity is contingent upon the correct implementation of the pre-processing, the tuning and ensuring clear separation of the participants. The number of independent test participants, uncertainty estimation with paired samples and proper treatment of repeated seeds are all important to ensure statistical validity. The cohort composition and the scanner protocols of the study limit external validity, as does the use of a research dataset. Definitions of the diagnostic label and the failure to match a binary endpoint with the larger statement of the classification of Alzheimer's disease are both sources of construct validity issues. For this revision, the following files are missing: source code, checkpoints, manifests split by subject, raw prediction files and training logs, which limits the reporting of the validity. Independent reproduction of the numerical results is not possible because of these limitations.

***F. Governance and clinical translation:*** Prior to clinical use, intended use should be carefully specified, including the target population, care setting, type of model output, user, and action to be taken if the output is positive or indeterminate. A retrospective research classification experiment is not a prospective benefit classification experiment. Clinical assessment should not only consider diagnostic discrimination but also assess workflow impact, false alert burden, calibration over time, failure modes, and interpretation by clinicians. The right governance also involves access control, privacy protection, versioning and deployment performance monitoring. The ADNI cohort is a cohort for research and it should be described according to the data-use and -acknowledgment requirements of the data that is downloaded. The manuscript should NOT be interpreted as an endorsement of the model or its results by ADNI. Additional validation and relevant institutional, regulatory and ethical review would need to be carried out before its proposed clinical use. These protection measures are of special relevance as structural changes with age, vascular disease, scanner differences, and comorbidities can impact the MRI appearance and model prediction.

***G. Recommended reporting checklist:*** When preparing an empirical paper for submission, include: (1) cohort flow diagram, and exact numbers for inclusion/exclusion; (2) participant-level train/validation/test manifests; (3) prespecified primary endpoint, and a rule for threshold selection; (4) per-class confusion matrices and ROC curves; (5) paired confidence intervals for differences between models; (6) reliability diagrams and ECE details; (7) per-seed metrics and variability; (8) hardware and timing protocol; (9) external validation, if available; and (10) code, environment files, and model-selection logs. If something is not done, make it clear as to why.

The experiment further differed in that H. was limited in some aspects and the next experiments were conducted.

In addition to the sample size limitations, the current evidence has a number of other limitations. First, the population described is a research population and may not reflect all patients with dementia seen in primary care, memory clinics or in under-represented communities. Second, the report provided does not make any conclusions about the number of visits to the study sites by participants or about how the longitudinal scans were managed. Third, the efficiency numbers are not hardware and timing specific, so the reported efficiency cannot be replicated exactly from the manuscript. Fourth, no codebase, prediction files and split manifest are provided, and confidence intervals and seed variability cannot be verified. Lastly, there is no external test cohort of independent data to quantify transportability between sites or scanners.

The next experiment should be able to reproduce the binary endpoint from a frozen participant manifest and release perseed prediction files. The second stage should include a clearly defined 3-class AD/MCI/CN task for evaluation for which class-wise and macro-averaged scores are calculated, not using the binary scores as evidence for the new evaluation. The third stage should test external generalization and calibration drift. Explain ability methods e.g. saliency maps, occlusion sensitivity should only be assessed after these steps, and explanations should be tested for stability and anatomical plausibility, but not as evidence of causal reasoning. The path to internal reproduction to potential clinical use is described in stages in Fig. 4.

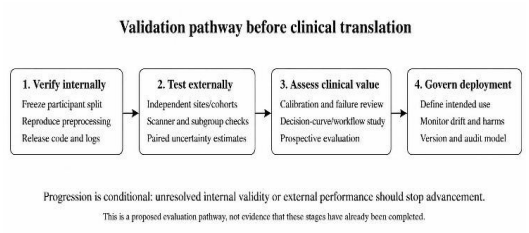



Fig. 4 Proposed internal reproduction to governed clinical deployment pathway. The stages are recommendations, and not completed study activities.

## VI. CONCLUSION

The manuscript presents an expanded benchmark framework for comparing CNNs and Vision Transformers in structural MRI-based AD classification. The source-reported results place Swin-Tiny at 85.1% balanced accuracy and DenseNet121 at 83.7% for the AD-versus-CN endpoint, without evidence in the supplied report of a statistically reliable difference. Efficiency measurements suggest that model selection involves a trade-off among discrimination, latency, memory, and calibration. The strongest contribution is therefore not a claim that one architecture is definitively superior, but a protocol emphasizing participant-level separation, uncertainty, calibration, operational reporting, and reproducibility. Before the paper is represented as a completed empirical study, the numerical results and dataset accounting must be verified from the actual experimental artifacts. Future work should prioritize external multi-site evaluation, clinically challenging MCI-related endpoints, MRI-specific pretraining, volumetric modeling, subgroup robustness, and explain ability.

## REFERENCES


[1]. P. Scheltens et al., “Alzheimer’s disease,” The Lancet, vol. 388, no. 10043, pp. 505–517, 2016, doi: 10.1016/S0140-6736(15)01124-1.

[2]. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in Proc. Advances in Neural Information Processing Systems, 2012, pp. 1097– 1105.

[3]. K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[4]. G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4700–4708, doi: 10.1109/CVPR.2017.243.

[5]. M. Tan and Q. V. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in Proc. Int. Conf. Machine Learning (ICML), 2019, pp. 6105–6114.

[6]. Dosovitskiy et al., “An image is worth 16×16 words: Transformers for image recognition at scale,” in Proc. Int. Conf. Learning Representations (ICLR), 2021.

[7]. Z. Liu et al., “Swin Transformer: Hierarchical vision transformer using shifted windows,” in Proc. IEEE/CVF Int. Conf. Computer Vision(ICCV), 2021, pp.1001210022,doi:10.1109/ICCV48922.2021.00986.

[8]. R. C. Petersen et al., “Alzheimer’s Disease Neuroimaging Initiative (ADNI): Clinical characterization,” Neurology, vol. 74, no. 3, pp. 201– 209, 2010, doi: 10.1212/WNL.0b013e3181cb3e25.

[9]. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, pp. 436–444, 2015, doi: 10.1038/nature14539.

[10]. S. M. Smith et al., "Advances in functional and structural MR image analysis and implementation as FSL," NeuroImage, vol. 23, suppl. 1, pp. S208–S219, 2004, doi: 10.1016/j.neuroimage.2004.07.051.

[11]. J. Ashburner and K. J. Friston, "Voxel-based morphometry—The methods," NeuroImage, vol. 11, no. 6, pp. 805–821, 2000, doi: 10.1006/nimg.2000.0582.

[12]. G. Litjens et al., "A survey on deep learning in medical image analysis," Medical Image Analysis, vol. 42, pp. 60–88, 2017, doi: 10.1016/j.media.2017.07.005.

[13]. Kumar, R. (2025, January). Artificial intelligence's metamorphosis in higher education (HEIs): A transformational study. Social Research Foundation, 9(1).

[14]. G. S. Collins et al., "TRIPOD-AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods," BMJ, vol. 385, 2024, e078378, doi: 10.1136/bmj2023-078378.

[15]. R. Jack Jr. et al., "The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods," Journal of Magnetic Resonance Imaging, vol. 27, no. 4, pp. 685–691, 2008, doi: 10.1002/jmri.21049.

# Biomechanical Analysis and Artificial Intelligence-Based Performance Assessment of Karate Punching Techniques: A Comprehensive Review

**[1] Anija, [2] Dr. Shalu Gupta**

[1] Research Scholar, Department of Computer Application, Guru Kashi University, Talwandi Sabo, Bathinda

[2]Assistant Professor, Department of Computer Application, Guru Kashi University, Talwandi Sabo, Bathinda.

Email Id: [1] anijamittal@gmail.com, [2] shalu2324@gmail.com

Accepted:20-09-2026 Published:12-10-2026 Page No. 71 – 78 DOI: 10.5281/zenodo.24391681

*Abstract - Karate punches are complex biomechanical movements that require neuromuscular control, balance, strength, speed, and coordination. The correct evaluation of these techniques is important to improve athletic performance, to find technical errors and to support effective training. Conventional biomechanical assessment methods, including motion capture systems, electromyography (EMG), inertial measurement units (IMUs) and video analysis, have contributed significantly to our understanding of punching velocity, acceleration, joint kinematics, impact force and muscle activation patterns. However, these techniques are usually based on specialized equipment, controlled experimental conditions and limited opportunities for automated real-time performance assessment. This review considers existing studies of biomechanical properties of different types of karate punches: Choku-Zuki, Oi-Zuki, Gyaku-Zuki, Kizami-Zuki, Mawashi-Zuki, Age-Zuki, as well as combination punches.It also discusses the use of artificial intelligence, computer vision, human pose estimation, and deep learning methods such as Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and Spatial-Temporal Graph Convolutional Networks (ST-GCN) for automatic recognition and performance evaluation of karate actions. Some previous studies suggest the possibility of sensor-based and vision-based methods for karate movement recognition and technical performance evaluation. However, challenges still remain in terms of dataset availability, complex movement recognition, movement variability and technique correctness assessment. The review emphasizes the research gap between biomechanical analysis and automated performance evaluation using artificial intelligence, and points to opportunities for the development of intelligent, vision-based systems to support objective karate punching analysis and personalized coaching.*

*Keywords-Karate punching techniques, Biomechanical analysis, Artificial intelligence, Deep learning, Human pose estimation, Computer vision, Action recognition, Performance assessment.*

## I. INTRODUCTION

Karate is one of the forms of martial arts that combines various methods of attacking and defending, which use strength, agility, balance, coordination, and body movements. In these methods, punching is one of those methods that has an important role in the practice of karate both in competition and in terms of technique. In order for a punching technique to be successful, it is not enough that just the upper limbs make the required movements; the lower limbs, trunk, shoulders, and arms need to work together in the process. Biomechanics offers a scientific method of evaluating these attributes using parameters such as the velocity of punching, acceleration, impact force, joint angles, duration of movement, and muscle activation. It has been proven through earlier research that punching skills are affected by the choice of technique, individual physical attributes, neuromuscular coordination, and the experience level of the practitioners [1], [2], [4].

## II. RELATED WORK

The biomechanics of various punching techniques (fist punches, karate chops, and palm strikes). Fifty participants (29 males, 21 females) with no specific martial arts training were included in the study. The punching velocity, maximum impact force, impulse, impact duration and effective mass were measured with a high-speed camera (2000 frames/s) and a force plate. Results showed that there were significant differences in the punching techniques with fist punches generally producing greater impulse and effective mass and palm strikes showed higher maximum forces and velocities. Punching performance was also affected by body weight, gender, hand dominance and individual ability. However, the study was limited to laboratory biomechanical measures and did not include the automated recognition of technique or performance assessment using artificial intelligence. The results provide a useful basis for identifying biomechanical parameters relevant in future computer vision and deep learning-based punching analysis [1]. The association between the maximum power of the lower limbs and the velocity of punching for various karate attacking movements with static and dynamic exercises. In the experiment, 10 karate athletes participated and executed Gyaku-Tsuki and the combined techniques of Kizami-Tsuki and Gyaku-Tsuki. To measure the power of the upper and lower limbs during bench press and back squat exercises, an inertial measurement unit was used. Punching speed and technique were analyzed through video analysis using Kinovea software. It was observed that there was a strong positive correlation between lower limb power and punching speed ($r = 0.66–0.80$). In regard to the Kizami-Tsuki and Gyaku-Tsuki combination technique, the speed of punches was faster compared to that observed in the Gyaku-Tsuki technique alone. Nonetheless, this study did not include automated performance analysis based on deep learning and involved only a few male athletes. Based on the results of this study, it can be stated that lower limb power and speed of punches are important factors in karate performance [2]. An innovative approach based on electromyography (EMG) to assess the efficiency of movements and muscle coordination in Gyaku-Tsuki (reverse punch). The experiment was conducted with one male karate champion, who executed 20 punches using 16 wireless EMG sensors with attached three-axis accelerometers, along with synchronized video footage. The five fastest punches in terms of forearm acceleration were taken into account. The novel method focused on muscle activation sequence and activation intensity visualization depending on muscle groups. It was found that muscles exhibited rapid and well-coordinated activation, with onset time varying between -0.308 and -0.110 seconds before punch completion. The results of the analysis revealed the significance of stabilization of the lower limbs, rotation of the torso, and coordination of the upper limbs in the process of efficient execution of punching. At the same time, the analysis was performed using only one athlete and only one punching style without using the techniques of automatic detection and deep learning for performance analysis [3]. Kinematics and electromyographic properties of Choku-Zuki technique in karate. This research included 37 subjects: 18 karate athletes and 19 people with no experience in karate at all. For data collection, electromagnetic motion tracking sensors and surface electromyography (EMG) were used in order to examine joint movements, angular velocity, acceleration, duration of movements, and muscle activity. It was found that there were distinctive differences between experienced subjects and subjects without experience in karate. Experienced athletes showed better coordination of movement, shorter joint movements range, and closer to impact peak of angular velocity and muscle activation. The findings brought out the need for coordination between the nervous system and muscles, as well as the need for proper sequence of proximal to distal

movement while throwing punches. Nevertheless, the research was conducted using Choku-Zuki punches only and did not analyze automation of technique recognition and artificial intelligence-based performance evaluation. The results will be important for future research [4]. An optical motion capture database that includes kinematic data of karate skills executed by Kyokushin karate practitioners who were either beginners or advanced. In the study, there were 37 subjects from various age groups and skill levels executing Gyaku-Zuki (reverse punch) and various kicks under three conditions: in the air, with a training shield, and with an opponent. The kinematics were acquired using a Vicon optical motion capture system working at 250Hz with 39 reflective markers on the body segments. The total number of recordings made was 1,411 with 3,229 kick and punch executions. The data captured made possible the analyses of joint angles, motion trajectories, and kinematic differences among the different techniques and execution conditions. This paper showed how motion capture data can be used in biomechanics and machine learning techniques for recognizing karate techniques. But the paper mainly concentrated on developing the dataset and did not go further to build and test a deep learning algorithm for automatic punch recognition and performance evaluation [5]. An automated karate movement recognition and performance evaluation system named iKarate. The iKarate system is designed to assist practitioners and coaches in identifying incorrect karate techniques. Three-dimensional skeletal joint coordinates were recorded by a Microsoft Kinect v2 sensor and Z-score normalization was applied to reduce variations due to different body proportions. Fast Dynamic Time Warping (F-DTW) was applied to compare movement sequences executed with different speeds and to classify techniques based on similarity to reference movements. The study examined four karate techniques, Age-Uke, Mae-Geri, Gedan-Barai and Soto-Uke. The result of the experiment showed an average recognition accuracy of 90%, 100% for Age-Uke and Mae-Geri, and 80% for Gedan-Barai and Soto-Uke. The system also provided instant feedback on the performance including correctness of the movement and suggestions for improvement. However, the approach was based on Kinect-based skeletal tracking and traditional sequence-matching algorithm rather than deep learning, which suggests that there are opportunities for more sophisticated vision-based karate action recognition systems [6]. A deep learning approach for automatic recognition and kinematic analysis of karate punching techniques using wearable acceleration sensors and Convolutional Neural Networks (CNNs) Sixteen experienced karate practitioners performed 1,912 punching movements (Yun Tsuki, Mawashi Tsuki, Age Tsuki and Uraken) and a non-punch movement class. IMUs were placed on both wrists to collect three-axis linear acceleration data. The collected data were used to train and compare different architectures: Multilayer Perceptron (MLP), one-dimensional CNN and two-dimensional CNN. The experimental results showed that the most accurate 2D CNN classifier provided a multiclass classification accuracy of 96% and an average F1-score of 0.97. In addition, the kinematic analysis allowed to detect discrepancies in punching velocity and movement properties of the left and right hand. The research confirmed the high efficiency of deep learning for punch recognition using sensors. At the same time, the system was based on the use of wearable sensors and did not include marker less video pose estimation and assessment of technique correctness [7]. The researchers have introduced a deep learning approach to automatic recognition of karate kata types and assessing the correctness of movements based on the hybrid CNN-LSTM architecture. In the process, 100 videos demonstrating four types of karate kata done by seven skilled karateka have been used. Using

MediaPipe BlazePose, 33 body landmarks have been detected in each video frame, and then they have been transformed into 132-dimensional feature vectors. The researchers compared the performance of CNN and hybrid CNN-LSTM architectures for the purpose of karate kata classification and correct movements evaluation. The experimental results showed that the CNN-LSTM model had an accuracy rate of 95.74%, better than the CNN model that had an accuracy rate of 91.49%. These results indicated the importance of using the features of space and time in detecting complicated karate movements and correctness in executing them. But the experiment suffered from using a small number of data, controlled environment of recording, and only four classes of kata [8]. A smart coaching system through deep learning and computer vision for automated recognition, classification, and evaluation of Shotokan Karate movements. The authors created a Multiview dataset based on eight basic karate moves carried out by a sixth-DAN black belt and captured in the frontal view, left and right cameras' views. The algorithms of OpenPose and FastPose were used for extracting keypoints of the body skeleton. LSTM and ST-GCN models were used for recognition and classification of the movements. Also, an attempt was made to use VIBE algorithm for 3D visualization of the movement. The results from experiments revealed that the recognition accuracy rate of LSTM was 96%, better than that of ST-GCN, whose recognition accuracy rate was 91.01%. The results showed that the use of multi-view pose estimation and deep learning was effective in recognizing karate movements and smart coaching. But there were still challenges with the stability of ST-GCN and hand gesture recognition of VIBE [9].

Beyond karate-specific studies, research on human pose estimation and action recognition provides methods that can support automated technique analysis. Temporal consistency approaches address variations in pose estimates across video frames [10], while surveys of vision-based pose estimation describe deep-learning methods for extracting human body keypoints [11]. Beyond karate-specific studies, research on human pose estimation and action recognition provides methods that can support automated technique analysis. Temporal consistency approaches address variations in pose estimates across video frames [10], while surveys of vision-based pose estimation describe deep-learning methods for extracting human body keypoints [11].

Pose-feature fusion with CNN-LSTM models has also been investigated for multiview human action recognition [18], and spatio-temporal graph convolutional networks have been applied to skeleton-based action analysis in vision-based fall detection [19]. Although these studies are not direct evaluations of karate punches, their approaches are relevant to designing systems that model the spatial structure and temporal evolution of punching movements.

Other related work contributes complementary ideas for movement representation and sensing. Limb synchronization has been proposed as an indicator of movement quality in karate [13], and motion-capture recordings can be averaged to generate movement templates for comparison [16]. Multi-sensor information fusion offers a broader framework for combining heterogeneous activity data [14], while hand-gesture recognition research illustrates the use of sensor signals and machine-learning algorithms for identifying hand movements [12]. Image-compression research is relevant to the practical handling and storage of image data in vision-based pipelines, although it does not directly evaluate karate recognition [15]. Likewise, neural-network-based workload optimization provides a general example of data-driven support for training-related decisions, but it should not be interpreted as evidence of punch recognition or biomechanical assessment [17].

## III. CATEGORIES OF KARATE PUNCHING TECHNIQUES

Punching techniques in karate differ depending on the direction of motion, hand used for punching, position of the body, position of the target, and sequence of the technique performance. Some categories of punching techniques include straight punch, step punch, back punch, circular punch, rising punch, and combined punch. Different categories of punching techniques have unique biomechanical properties.

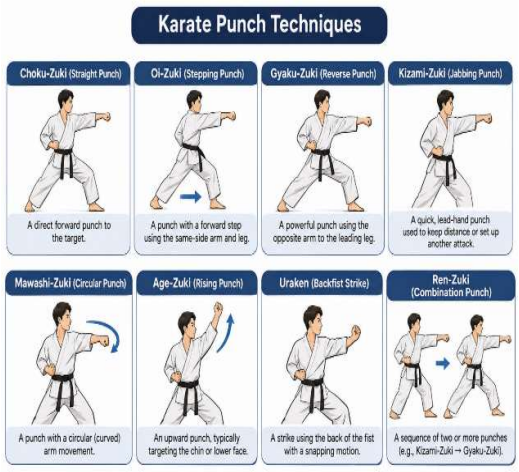



Fig. 1. SHOWS DIFFERENT CATEGORIES OF KARATE PUNCHING TECHNIQUES

***A. Straight Punches (Choku-Zuki):*** The Choku-Zuki punching technique is one of the most basic forms of punching techniques in karate. It is defined by the straight-out forward extension of the arm to the target. This technique entails coordination of the shoulder, elbow, and wrist joints along with rotation of the trunk and muscle contraction. Punching speed, joint angle displacement, acceleration, time taken to execute the movement, and muscle coordination are some of the parameters that can help analyze this technique [3].

***B. Stepping Punches (Oi-Zuki):*** The Oi-Zuki, which is another name for lunge or stepping punch, consists of a combination of moving the body forward as well as punching using the hand that is located on the same side as the foot that moves forward. Coordination of the movement of the foot, movement of the body, stability of posture, and arm motion are required. Timing of the motion of the lower and upper parts of the body and postural stability during the performance of the technique will determine its effectiveness.

***C. Reverse Punches (Gyaku-Zuki):*** The execution of Gyaku-Zuki is done by means of the limb that is not leading, and it is considered to be one of the most studied punching techniques in karate biomechanics. This technique requires the coordination of rotation of the hips, body, extension of the limbs and muscle contraction. The connection between the maximal power of the lower limbs and the speed of punches, establishing that there was a positive correlation between these variables [2], [3].

***D. Jabbing Punches (Kizami-Zuki):*** Kizami-Zuki is a punching technique using the lead hand for delivering quick attacks and keeping an effective fighting position. This technique is different from the reverse punch because it utilizes the lead arm more, and focuses on speed, timing, and body coordination. The biomechanics of Kizami-Zuki may consist of punching velocity, shoulder and elbow movements, timing, and coordination with the lower body parts. Kizami-Tsuki combined with Gyaku-Tsuki and found higher punching velocities when combining the techniques compared to executing Gyaku-Tsuki alone [2].

***E. Circular Punches (Mawashi-Zuki):*** The Mawashi-Zuki punch is defined by its characteristic curved path rather than a straight-line trajectory. The punch execution is composed of the coordination of arm, shoulder, and body movements. Curved movement generates unique acceleration and velocity characteristics that differ from those of straight-line punches. The automatic recognition of karate punches such as Mawashi Tsuki based on acceleration data collected by wrist-worn sensors in combination with convolutional neural networks [7].

***F. Rising Punches (Age-Zuki):***

Age-Zuki is the punching technique where the punching is done in an upwards direction, with control of the elbow, shoulder muscles, and proper fist position. This punching technique has a different movement trajectory and acceleration pattern compared to other punches such as linear and circular punches. Tsuki into their framework for recognizing punches using deep learning [7].

***G. Backfist Strikes (Uraken):*** Uraken is an unusual technique performed through the use of the back of the hand, and it is frequently related to quick arm motions. Despite being a type of backfist strike and not a traditional punch, the technique is pertinent to the topic of automatic karate actions detection due to its unique movement path and acceleration pattern. According to Uraken was one of the techniques studied by convolutional neural networks.

***H. Combination Punches (Ren-Zuki and Kizami-Zuki–Gyaku-Zuki):*** Combination punching is a process where two or more punches are performed sequentially, entailing proper timing and fast transitions of the movement. Combination punching entails coordination of both upper and lower body parts to produce enough speed and stability during the technique. Importance of leg power in combination punching has been proven. Analysis of combination punches using automation technology entails identification of individual movements, as well as timing and transitions [2].

***TABLE I. COMPARISON OF REPRESENTATIVE STUDIES***

| Ref. No | Sample | Method | Focus | Major Findings |
|---|---|---|---|---|
| [1] | Fifty participants (29 males, 21 females) | High-speed videography and force plate-based biomechanical analysis | Punching velocity, impact force, impulse, impact duration, and effective mass | Found major differences in the biomechanics of fist punches, karate chops and palm strikes. Demonstrated the significance of punch type and individual characteristics on striking performance and offered measurable parameters pertinent to future automated performance assessment. |
| [2] | 10 national-level male karate athletes performing individual and combined punching techniques under static and dynamic conditions. | Biomechanical experimentation using the inertial measurement unit, the bench press and back squat power tests, and video kinematics analysis through the Kinovea software application. | Relationship between lower-limb maximal power and punching speed during karate attacking techniques | Detected a high correlation between the maximal power of the lower limb and punching velocity (r = 0.66 – 0.80). It was shown that combination techniques were characterized by higher velocities than the isolated application of Gyaku-Tsuki. |
| [3] | One elite male karate athlete performing 20 Gyaku-Tsuki punches. | Analysis of surface EMG using 16 wireless sensors, 3D accelerometers, video recording, and Python-based signal processing. Analysis of surface EMG using 16 wireless sensors, 3D accelerometers, video recording, and Python-based signal processing. | Muscle activation sequences, neuromuscular coordination, and movement efficiency during Gyaku-Tsuki execution. | Discovered patterns of coordination among muscles of lower limbs, trunk and upper limbs. Shown the feasibility of using EMG based analysis to detect technical mistakes and optimize punching technique in karate. |
| [4] | 37 participants: 18 experienced karate athletes and 19 non-karate participants performing Choku-Zuki. | Electromagnetic motion tracking and surface EMG analysis of joint kinematics and muscle activation. | Joint movement, angular velocity, acceleration, movement timing, and neuromuscular coordination during Choku-Zuki. | Experienced athletes showed reduced movement amplitude, improved coordination, and increased peak angular velocity nearer impact. The study provided biomechanical parameters for the objective evaluation of karate punching performance. |
| [5] | 37 beginner and advanced Kyokushin karate athletes | Vicon optical motion capture (250 Hz), 39 reflective markers, and 3D kinematic data collection. | Development of a karate motion capture dataset for analyzing joint movements, trajectories, and technique execution. | Created a database comprising 1,411 audio clips, which include 3,229 kicks and punches. Suitable for motion analysis and machine learning technique recognition. |
| [6] | Four karate techniques (Age-Uke, Mae-Geri, Gedan-Barai and Soto-Uke); real-time testing with five karate players | Kinect v2 sensor, 3D skeletal joint tracking, Z-score normalization and Fast Dynamic Time Warping (F-DTW) | The study examined four karate techniques, Age-Uke, Mae-Geri, Gedan-Barai and Soto-Uke. | Achieved average recognition accuracy of 90 %. The system could differentiate between right and wrong moves, and offer feedback to improve technique. |
| [7] | 6 experienced karate athletes (12 male, 4 female); 1,912 movements involving Yun Tsuki, Mawashi Tsuki, Age Tsuki, Uraken and a non-punch class | Wrist-mounted acceleration sensors (IMUs), Multilayer Perceptron (MLP), 1D-CNN and 2D-CNN | Automatic recognition and classification of karate punches using deep learning and kinematic analysis | The highest performing 2D-CNN had a classification accuracy of 96% and average F1-score of 0.97. The research showed successful sensing-based punching identification, and also the different kinematic features of punching with left versus right hands were revealed. |
| [8] | 100 video clips of four karate kata classes | MediaPipe BlazePose for extracting 33 body keypoints; CNN and hybrid CNN–LSTM models | Automatic karate kata classification and movement correctness evaluation using deep learning | The CNN–LSTM model achieved 95.74% classification accuracy, outperforming the CNN-only model (91.49%). Temporal feature learning improved kata recognition and movement correctness assessment. |
| [9] | Eight Shotokan Karate techniques performed by a sixth-DAN black belt expert, captured from three camera views (front, left and right) | OpenPose, FastPose, LSTM, ST-GCN and VIBE-based 3D visualization | Automatic karate action recognition, movement classification, performance scoring and intelligent coaching | In addition, the LSTM model performed at 96% accuracy, surpassing ST-GCN, which recorded 91.01%. Multiview analysis of skeletal data proved to be an efficient method of classifying karate moves and performance evaluation; however, there were some shortcomings in hand gesture recognition and ST-GCN instability. |

## EVALUATION METRICS FOR BIOMECHANICAL ANALYSIS AND AI-BASED PERFORMANCE ASSESSMENT

These formulas can be incorporated into the "Evaluation Metrics" part of your review paper, since they include biomechanical performance, artificial intelligence action recognition, pose estimation, skill assessment, and computation speed. It is strongly advised to use these formulas as evaluation metrics; they must not be presented as results of experiments unless relevant data were collected and analyzed.

### *A. Biomechanical performance metrics*

1.Punching velocity

Punching velocity measures how quickly the first moves during the execution of a punch.

$$v = \frac{d_2 - d_1}{t_2 - t_1}$$

Where:

d_ (1,) d_2= Initial and Final first positions.

t_ (1,) t_2= Corresponding time instances.

For three- dimensional movement, velocity can be calculated as [1]:

$$v = \sqrt{v_x^2 + v_y^2 + v_z^2}$$

2. Acceleration

Angular velocity measures the speed of rotation of a joint, such as the shoulder, elbow, or wrist.

$$\alpha = \frac{v_2 - v_1}{t_2 - t_1}$$

Where v_1 and v_2 represent intital and final velocities. [1]

3. Angular velocity

Angular velocity measures the speed of rotation of a joint, such as the shoulder, elbow, or wrist.

$$\omega = \frac{\theta_2 - \theta_1}{t_2 - t_1}$$

Where:

- ⟦θ_(1, ) θ⟧ _2 = joint angles at two-time instances.
- t_(1, ) t_2 = corresponding time instances.

[2]

***B. AI-based action recognition metrics:*** They assess the accuracy of a deep learning system in distinguishing various punching styles of karate.

Accuracy

$$Accuracy = \frac{TP+T}{TP+TN+FP+FN}$$

Precision

$$Precision = \frac{TP}{TP + FP}$$

Recall

$$Recall = \frac{TP}{TP+FN} \times 100$$

F1-Score

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Where:

- TP= True Positive
- TN= True Negative
- FP= False Positive
- FN= False Negative

In case of the multi-class karate punch recognition problem, the precision, recall, and F1-scores need to be provided using macro or weighted averaging [6][7].

***C. Human pose estimation metrics:*** These metrics will be helpful if Open Pose or any other pose estimation algorithm is used for detecting body key points while performing karate punches.

*1. Percentage of Correct Keypoints (PCK):* This metric determines the percentage of correctly detected body keypoints at a given distance threshold.

$$PCK(\alpha) = \frac{1}{NK} \sum_{i=1}^{N} \cdot \sum_{j=1}^{K} 1\left(\frac{||p_{ij} - \hat{p}_{ij}||2}{S_i} < \alpha\right) \times 100$$

*2. Mean Per Joint Position Error (MPJPE):* MPJPE measures the average Euclidean distance between predicted and ground-truth 3D joint positions.

$$\text{MPJPE} = = \frac{1}{NK} \sum_{i=1}^{N} \cdot \sum_{j=1}^{K} 1 \, ||p_{ij} - \hat{p}_{ij}||2$$

*3. Object Keypoint Similarity (OKS):* OKS measures the similarity between predicted and ground-truth body keypoints.

$$OKS = \frac{\sum_i exp\left(\frac{-d^2}{2S^2 k_i^2}\right) v_i}{\sum_i v_i}$$

*4.Keypoint Localization Error (KLE):* KLE measures the average distance between actual and predicted positions of individual body landmarks.

$$KLE = \frac{1}{NK} \sum_{i=1}^{N} \sqrt{(\chi_i - \hat{\chi}_i)^2 + (y_i - \hat{y}_i)^2}$$

*5. Mean Average Precision (mAP):* mAP evaluates overall pose detection performance across different OKS thresholds. The wider literature also points to the importance of robust pose estimation, multimodal data integration, and patio-temporal modelling [10], [11], [14].

$$MAP = \frac{1}{T} \sum_{t=1}^{T} AP_{OKS = t}$$

## II. RESEARCH GAP

Movement synchronization, motion-template comparison, sensor fusion, and spatio-temporal learning offer complementary directions for future systems [18], [19]. There are several gaps in karate punching research, such as the limited availability of punch-specific datasets, the insufficient coverage of different techniques and execution conditions, and the reliance on specialized sensors. “Existing AI models are more focused on action recognition

rather than biomechanical performance and accuracy assessment. In addition, practical applications are limited by poor integration of spatial and temporal features, inconsistencies in the evaluation methods, and lack of personalized feedback. Future work should focus on the development of an integrated explainable deep learning framework that employs marker less pose estimation and biomechanical features to provide accurate, reliable, and personalized performance assessment.

## VI. CONCLUSION

This review explored the biomechanical characteristics of karate punching techniques and the use of artificial intelligence, computer vision, pose estimation and deep learning for their recognition and performance evaluation. The literature emphasizes the significance of movement coordination, punching velocity, joint angles, muscle activation and other biomechanical parameters, and demonstrates the potential of sensor-based and pose-based deep-learning models for action recognition. However, there are limitations in the punch-specific datasets, athlete diversity, real-world validation, and standardized evaluation. Future research should combine marker less pose estimation, spatio-temporal deep learning and biomechanical indicators for accurate technique assessment, identification of execution errors and personalized coaching feedback.

## REFERENCES


[1] J. Adamec, P. Hofer, S. Pittner, F. Monticelli, M. Graw, and J. Schöpfer, "Biomechanical assessment of various punching techniques," International Journal of Legal Medicine, vol. 135, no. 3, pp. 853–859, 2021.

[2] F. Quinzi, G. Rosellini, and P. Sbriccoli, "Lower limb maximal power predicts punching speed in different static and dynamic attacking techniques in karate," The Journal of Strength & Conditioning Research, vol. 36, no. 5, pp. 1353–1359, 2022.

[3] L. Csákvári, B. Kopper, and T. Horváth, "A novel, sport-specific EMG-based method to evaluate movement efficiency in karate punching," Sports, vol. 13, no. 7, Art. no. 218, 2025.

[4] A. M. VencesBrito, M. A. R. Ferreira, N. Cortes, O. Fernandes, and P. Pezarat-Correia, "Kinematic and electromyographic analyses of a karate punch," Journal of Electromyography and Kinesiology, vol. 21, no. 6, pp. 1023–1029, 2011.

[5] A. Szczęsna, M. Błaszczyszyn, and M. Pawlyta, "Optical motion capture dataset of selected techniques in beginner and advanced Kyokushin karate athletes," Scientific Data, vol. 8, no. 1, Art. no. 13, 2021.

[6] B. Emad, O. Atef, Y. Shams, A. El-Kerdany, N. Shorim, A. Nabil, and A. Atia, "iKarate: Improving karate kata," Procedia Computer Science, vol. 170, pp. 466–473, 2020.

[7] A. Labintsev, I. Khasanshin, D. Balashov, M. Bocharov, and K. Bublikov, "Recognition punches in karate using acceleration sensors and convolution neural networks," IEEE Access, vol. 9, pp. 138106–138119, 2021.

[8] N. Abdulrahman, Z. Zainuddin, and I. Nurtanio, "A deep learning-based framework using CNN+LSTM for karate kata classification and correctness evaluation," Engineering, Technology & Applied Science Research, vol. 16, no. 1, pp. 32619–32624, 2026.

[9] A. Aaroud, "Applying deep learning and computer vision techniques for an e-sport and smart coaching system using a multiview dataset: Case of Shotokan karate," International Journal of Online and Biomedical Engineering (iJOE).

[10] Y. Li, K. Li, X. Wang, and R. Y. Da Xu, "Exploring temporal consistency for human pose estimation in videos," Pattern Recognition, vol. 103, Art. no. 107258, 2020.

[11] G. Lan, Y. Wu, F. Hu, and Q. Hao, "Vision-based human pose estimation via deep learning: A survey," IEEE Transactions on Human-Machine Systems, vol. 53, no. 1, pp. 253–268, 2022.

[12] A. K. Panda, R. Chakravarty, and S. Moulik, "Hand gesture recognition using flex sensor and machine learning algorithms," in Proc. 2020 IEEE-EMBS Conf. Biomed. Eng. Sci. (IECBES), Mar. 2021, pp. 449–453.

[13] P. Alborno, N. De Giorgis, A. Camurri, and E. Puppo, "Limbs synchronisation as a measure of movement quality in karate," in Proc. 4th Int. Conf. Movement Computing, Jun. 2017, pp. 1–6.

[14] S. Qiu et al., "Multi-sensor information fusion based on machine learning for real applications in human activity recognition: State-of-the-art and research challenges," Information Fusion, vol. 80, pp. 241–265, 2022.

[15] H. T. Sadeeq, T. H. Hameed, A. S. Abdi, and A. N. Abdulfatah, "Image compression using neural networks: A review," International Journal of Online and Biomedical Engineering, vol. 17, no. 14, pp. 135–153, 2021.

# GKU JOURNAL OF MULTIDISCIPLINARY RESEARCH (GKUJMR)

- Exploring Diverse Perspective
- Advancing Knowledge
- Impacting Society

GKU Journal of Multidisciplinary Research (GKUJMR) is a peer-reviewed, open access journal that publish high quality research across a wide range of disciplines. The Journal aims to foster interdisciplinary dialogue, promote innovative thinking, and contribute to the advancement of knowledge for the benefit of the society.



**Paper Indexing In**

Google Scholar | OpenAIRE

## Contact Us


Guru Kashi University,
Talwandi Sabo, Bathinda
Punjab, India. (151302)

gkujournal@gku.ac.in
https://journal.gku.ac.in
+91 94644 79497


OPEN ACCESS
Knowledge for All