BenchCouncil Transactions
on Benchmarks, Standards and Evaluations

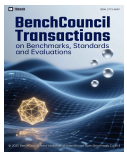
TBench
Volume 6, Issue 2
2026

Research Article

ISSN: 2772-4859

BenchCouncil Transactions on Benchmarks, Standards and Evaluations (TBench) is an open-access multi-disciplinary journal dedicated to benchmarks, standards, evaluations, optimizations, and data sets. This journal is a peer-reviewed, subsidized open access journal where The International Open Benchmark Council pays the OA fee. Authors do not have to pay any open access publication fee. However, at least one of the authors must register BenchCouncil International Symposium on Benchmarking, Measuring and Optimizing (Bench) (https://www.benchcouncil.org/bench/) and present their work. It seeks a fast-track publication with an average turnaround time of one month.

Contents



BenchCouncil Transactions on Benchmarks, Standards and Evaluations, 2026

DOI: https://doi.org/10.66834/b5qtx140
Research Article

RESEARCH ARTICLE

A Hybrid MCDM Framework for Assessing Financial Resilience and Trend Dynamics in Indian Commercial Banks

Priya Das[1,*] and Subir Kumar Sen[2]

[1]Research Scholar, Department of Commerce, Tripura University, Agartala, 799022, India and [2]Professor, Department of Commerce, Tripura University, Agartala, 799022, India
[*]Corresponding author. priya568das@gmail.com
Received on 1 January 2026; Accepted on 20 May 2026

Abstract

Assessing financial resilience in the banking sector requires an integrated framework that captures both cross-sectional strength and long-term resilience dynamics. In this study, financial resilience of Indian commercial banks during the period 2013–2024 is assessed by using a hybrid MCDM and non-parametric trend analysis approach. Eleven financial indicators covering solvency, asset quality, efficiency, and profitability are included in a composite resilience framework, and criterion weights were objectively determined using the MEREC method. The technique RAM is used to calculate annual composite resilience scores and ranks the 29 commercial banks. A Mann–Kendall time-series analysis is also applied to the final RAM scores to analyze long-term monotonic trends in bank-level and sector-wide resilience. The results showed that RAM scores are tightly clustered across banks, suggesting structural convergence in resilience levels. However, Kruskal-Wallis non-parametric test showed statistically significant differences in the banks' relative financial resilience across the study period. The MEREC-RAM ranking result showed Kotak Mahindra Bank Ltd. and Tamilnad Mercantile Bank Ltd. consistently appeared among top at the rankings. While, the Mann-kendall trend test revealed significant Improvement in the resilience of CSB Bank Ltd. and Bank of Maharashtra over the study period. Overall, combining year-wise relative rankings and monotonic resilience dynamics enables a comprehensive assessment of the stability of the Indian banking sector, which can offer key insights for regulators, policymakers, and bank management in strengthening the long-term financial resilience of the sector.

Key words: Banks, Financial Resilience, MCDM, MEREC, RAM, Mann-Kendall Trend Analysis

JEL codes: G21, G28, G32, C44, D81

1. Introduction

Financial resilience is the ability of an institution to manage, respond to, and bounce back from financial challenges such as economic downturns, reduced earnings, or unforeseen costs. For banks and other financial institutions, the financial resilience refers to the ability to withstand shocks, adapt to changing economic conditions, and maintain stability while continuing to meet obligations and support economic activities[1, 2]. In the context of emerging economies, where financial systems often face volatility, weak governance, and high exposure to non-performing assets (NPAs), evaluating the resilience of banks is particularly crucial[3, 4]. The resilience of financial systems is influenced by several factors, such as firm characteristics, capital adequacy, liquidity management, risk governance, and their ability to respond to external shocks[5]. Assessing financial resilience enables regulators and policymakers to identify vulnerabilities and design regulatory frameworks that enhance risk management and recovery mechanisms[2].

The Indian banking sector has experienced significant structural changes in the past decades. The heavily regulated regime of the 1980s constrained were limiting operational efficiency of Indian banks and their ability to respond to a rapidly expanding economy[6–8]. However, the economic reforms of the early 1990s, which included deregulation, partial privatisation of public sector banks, and interest rate liberalisation, brought improvements in productivity, competitiveness, and risk management practices[6, 9–11] . Conversely, the post-global financial crisis period highlighted persistent weaknesses in banking systems around the world, including India, in the form of huge NPAs and governance issues[12, 13]. The evolution of the banking sector in India has built upon previous structural reforms and has moved towards a more risk sensitive and resilient regulatory framework. This evolution has

been guided by the adoption of Basel I and II capital adequacy standards, the introduction of risk-based supervision, and governance reforms in public sector banks aimed at enhancing asset quality, transparency, and prudential discipline[14]. These efforts laid the groundwork for more specific post-crisis reforms, such as the Asset Quality Review (AQR) and the Insolvency and Bankruptcy Code (IBC) in 2015–2016 that enhanced balance-sheet recognition and codified mechanisms for stress resolution[15, 16]. Although greater regulation has led to better governance and operational efficiency, there is some evidence that heavier compliance requirements can reduce profitability and internal capital generation, particularly in times of economic stress[17]. In 2018, the Punjab National Bank (PNB) scam revealed serious gaps in internal controls, governance mechanisms, and risk oversight in Indian banks. It demonstrated the need for stronger processes, transparency, and institutional resilience to avoid systemic vulnerabilities [18]. At the same time, the digitalization of banking services has brought new operational risks[19].

These developments point to some of the ongoing challenges faced by the Indian banking sector in sustaining financial resilience throughout cyclical stress, rising NPAs, and balance-sheet pressures. Commercial banks, particularly public sector banks, which account for a substantial share of total banking assets, have remained vulnerable to capital adequacy constraints, asset quality deterioration, and volatility in profitability. Such vulnerabilities underline the importance of continuously monitoring the stability of banks and their capacity to absorb shocks arising from macroeconomic fluctuations and structural adjustments.

Therefore, banking sector resilience should be systematically and multidimensionally assessed to explore how banks cope with financial stress and adapt to regulatory and economic changes. In this context, this study aims to evaluate the financial resilience of Indian commercial banks over 12 years (2013–2024) using a multi-criteria decision-making (MCDM) framework that incorporates the Method based on the Removal Effects of Criteria (MEREC)-based objective weighting and the Root Assessment Method (RAM), supported by non-parametric trend analysis. The Method RAM is relatively new and first applied in measuring banking sector performance. The MEREC-RAM approach can be used to constructs composite resilience scores based on deviations from reference performance levels. It allows consistent aggregation of heterogeneous financial indicators without excessive sensitivity to scaling assumptions[20, 21]. This framework can be combined with the Mann–Kendall trend test to compare the relative resilience rankings and their evolution over time, extending MCDM applications beyond static performance comparisons. The unique methodology of the current study is the integration of the impact-driven weighting scheme with a transparent composite scoring approach and a dynamic time-series evaluation that does not rely on the subjective judgments or correlation-based weights used in other studies. Instead this model built on a method that assigns weights by evaluating the contribution of each indicator to the overall system performance[22].

Accordingly, the present study is organized into five major sections. Section 1 presents a brief introduction of the study; Section 2 outlines the background of the study, mentioning the relevant literature, as well as hypotheses development based on relevant theories. Section 3 presents the methodological steps and procedures, Section 4 discusses the major findings using tables and figures, and finally, Section 5 ends with the study conclusions mentioning managerial and policy implications, and future research directions.

2. Background of the Study

2.1. Financial Resilience Background

Financial resilience has become a key determinant in understanding how individuals, households, institutions, and economies handle and recover from financial shocks. It generally refers to the ability to maintain stability and well-being during disruptions by combining financial resources, skills, and institutional support. This idea encompasses proactive strategies, including accumulating savings, diversifying income sources, and utilizing effective financial tools, as well as having the capacity to recover or even advance to a better financial position[23]. Salignac et al.[1] proposed a foundational framework that explored the concept of financial resilience at the household and community levels in Australia. The study demonstrated the role that resources, skills, and access to opportunities play in resilience. This work was further extended to a global scale by Klapper and Lusardi[24], who showed that financial literacy is a significant predictor of financial resilience at both the individual and household levels, as well as over time[25]. Hamid et al.[4] investigated the determinants of financial resilience in the context of an emerging economy in Malaysia, highlighting the key role that savings, income diversification, and access to credit play in enabling individuals and households to build financial resilience. Salignac et al.[2] extended this to developing economies, where institutional and policy supports for national financial resilience were found to contribute to inclusive economic development. Jansson[26] linked financial resilience to firms, noting that financial position, profitability, and ownership structures impact a firm's ability to withstand downturns. Sreenivasan and Suresh[27] examined start-ups and found that good liquidity management, innovation, and preparedness help new entrepreneurs to manage change. In a study of U.S. households during the COVID-19 pandemic, Clark and Mitchell[28] found that fiscal assistance programs and savings buffers were important for individual financial resilience.

Financial resilience in banking and other financial institutions has been studied extensively, including international banks during the global financial crisis[29], banking systems around the world[30], the U.S. banking system[31], and a loan-level analysis of financial institutions in Mexico[32]. Alam et al. [33] conducted research on corporate resilience among firms in Bangladesh, specifically the relationship between corporate ownership patterns and the resilience of firms, Daadmehr[34] proposed a composite financial resilience index for workplaces and firms, which combines the multi-dimensional matrices (liquidity, leverage, and sustainability practices). Chen and Sun[35] presented an econometric approach to measuring financial resilience across institutions and economies that combine dynamic panel modelling and stress-testing indicators to capture the temporal response of financial entities to shocks.

2.2. MCDM in Bank Performance Evaluation

MCDM techniques have been extensively used for assessing bank performance in different institutional and regional contexts[36]. Most empirical studies have grouped MCDM approaches into criteria weighting methods and outranking methods. Studies have combined the two types of MCDM tools to get a robust performance ranking. AHP–TOPSIS and IV–TOPSIS was used to evaluate listed private banks in India,

which consistently showed the superiority of HDFC Bank[37], and a CRITIC-based TOPSIS framework was incorporated to evaluate public sector banks in India, which showed persistent performance differences[38]. Such hybrid frameworks have been used internationally. Nguyen et al. [39] and Yazdi et al. [40] showed that bank performance rankings during the COVID-19 period are contingent on the weighting and aggregation methods, and that objective weighting schemes are important.

Various alternative weighting and ranking methods have been compared to improve methodological robustness (e.g., Ünlü et al. [41], Ünvan & Ergenç [42], Wanke et al. [43], Sama et al. [44]). Recent contributions have examined hybrid and integrated models that combine subjective and objective factors to increase discrimination power and ranking stability[45–47]. Mohan and Irfan[8] also explained how artificial intelligence with MCDM tools was used to assess the performance of banks in India.

The literature confirms the effectiveness of MCDM frameworks in assessing bank performance; however, most studies remain confined to static, cross-sectional performance rankings. Limited attention has been paid to integrating MCDM-based composite indices with dynamic analyses that capture performance evolution over time, particularly in the context of financial resilience. This gap motivates the present study's use of an objective weighting scheme and a composite assessment framework that can support both ranking and temporal trend analysis.

2.3. Hypotheses Development

Financial resilience can be conceptualized as a multi-dimensional construct reflecting a bank's ability to absorb shocks, adapt to adverse conditions, and maintain core functions over time[48, 49]. While the theory of resilience posits that a bank's solvency strength, asset quality, operational efficiency, and profitability will result in resilience. However, the theory does not suggest that resilience will improve in a linear or monotonic way, especially in banking systems subject to regulatory reform, economic cycles, and changing risk profiles.

From a regulatory perspective, the Capital Buffer Theory[50, 51], as formalized by the Basel framework, states that capital adequacy and provisioning requirements are first and foremost intended to maintain minimum stability thresholds rather than to continually improve performance. Stricter asset classification standards, supervisory reviews, and corrective action frameworks are some of the regulatory interventions that force banks to rebalance their capital structures and portfolios[52]. These measures may increase system-wide stability, but it is uncertain how they will ultimately affect composite resilience measures because improvements in capital or asset quality may put short-term pressure on profitability and efficiency.

Moreover, Dynamic Capability Theory[53, 54] provides a strategic view, which states that resilience can only be improved over the long term if a bank has the ability to perceive upcoming threats and reorganize its resources in the face of evolving circumstances[54]. Due to differences in ownership structures, managerial skills, and strategic adaptability, not all banks tend to react to environmental change or regulatory compliance with higher overall resilience. Consequently, banks may exhibit heterogeneous and non-uniform resilience trajectories over time. Combined, these theoretical perspectives suggest that, at the aggregate level, bank financial resilience may not follow a systematic or statistically discernible long-term trend. Accordingly, the following null hypothesis is formulated:

H_0^1: There is no statistically significant monotonic trend in the composite RAM-based financial resilience scores of Indian commercial banks over the study period.

H_0^2: There is no statistically significant difference between private and public sector banks in India over the study period.

3. Methodology

This study examines the financial resilience of selected commercial banks in India from 2013-2014 to 2024-2025.The study employs an integrated MEREC-based RAM approach within an MCDM framework to determine the weights of criteria, rank banks based on their financial resilience composite scores, and then applies the Mann-Kendall trend analysis to identify the performance dynamics and resilience trends of banks over the years. It also employed Welch two-sample t-test and Kruskal–Wallis non-parametric test to examine bank specific and year-wise significant differences in resilience scores. The study used equal weights and entropy-based weighting to check for the sensitivity of weight change on the ranking outcomes. It further employed TOPSIS, RATMI, and MARCOS techniques to examine the robustness of the proposed MEREC-RAM model. Finally, the Kendall's tau coefficients are determined to analyse the concordance of ranking outcomes across the methods. Figure 1 presents the conceptual framework designed for conducting the research.

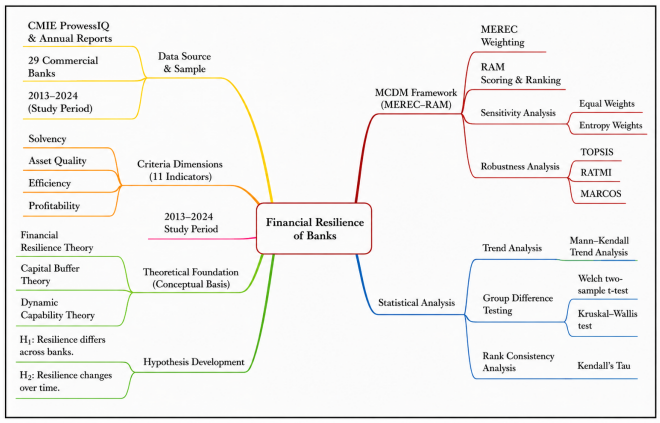


Figure 1. Conceptual Framework of the Study. (Source: Authors' Compilation)

Accordingly, the study integrated secondary data from the CMIE database and banks' annual reports from 2013 to 2024. The data collected for a sample of 29 commercial banks out of which 17 are private and 12 public sector counterparts. The study performed robust data cleaning, accounting for any missing data. Companies with continuous missing data for multiple variables were excluded from the analysis. To enhance the accuracy of the data, no approximation or rounding has been performed. Table 1 outlines the study sample, including bank codes.

Financial resilience in banks refers to the ability to absorb shocks, maintain core financial functions, and adapt to adverse economic and regulatory conditions. In line with the banking resilience and financial stability literature, resilience is conceptualised as a multidimensional construct encompassing solvency

Code	Name of Bank	Ownership	Code	Name of Bank	Ownership
CB1	Axis Bank Ltd.	**Private**	**CB16**	Indusind Bank Ltd.	**Private**
CB2	Bank of Baroda	**Public**	**CB17**	Jammu & Kashmir Bank Ltd.	**Private**
CB3	Bank of India	**Public**	**CB18**	Karnataka Bank Ltd.	**Private**
CB4	Bank of Maharashtra	**Public**	**CB19**	Karur Vysya Bank Ltd.	**Private**
CB5	CSB Bank Ltd.	**Private**	**CB20**	Kotak Mahindra Bank Ltd.	**Private**
CB6	Canara Bank	**Public**	**CB21**	Punjab & Sind Bank	**Public**
CB7	Central Bank of India	**Public**	**CB22**	Punjab National Bank	**Public**
CB8	City Union Bank Ltd.	**Private**	**CB23**	RBL Bank Ltd.	**Private**
CB9	DCB Bank Ltd.	**Private**	**CB24**	South Indian Bank Ltd.	**Private**
CB10	Dhanlaxmi Bank Ltd.	**Private**	**CB25**	State Bank of India	**Public**
CB11	Federal Bank Ltd.	**Private**	**CB26**	Tamilnad Mercantile Bank Ltd.	**Private**
CB12	HDFC Bank Ltd.	**Private**	**CB27**	UCO Bank	**Public**
CB13	ICICI Bank Ltd.	**Private**	**CB28**	Union Bank of India	**Public**
CB14	Indian Bank	**Public**	**CB29**	Yes Bank Ltd.	**Private**
CB15	Indian Overseas Bank	**Public**			

Table 1. Selected Commercial Banks (CB). (Source: Authors' Compilation)

Resilience	Code	Attributes	Descriptions	Expected Outcome
Solvency	C1	CAR	$\frac{Tier\ I\ Capital + Tier\ II\ Capital}{Risk - Weighted\ Assets}$	Max
	C2	Tier-1 CAP	$\frac{Tier\ I\ Capital}{Risk - Weighted\ Assets}$	Max
	C3	D/E	$\frac{Total\ debt}{Shareholders'\ Equity}$	Min
Assets Quality	C4	GNPA	$\frac{Gross\ NPAs}{Gross\ Advances}$	Min
	C5	NNPA	$\frac{Net\ NPAs}{Net\ Advances}$	Min
	C6	LGR	$\frac{Total\ Advances_t - Total\ Advances_{t-1}}{Total\ Advances_{t-1}}$	Min
	C7	PCR	$\frac{Total\ Provisions}{Gross\ NPAs}$	Max
Efficiency	C8	C-I	$\frac{Operating\ Expenses}{Total\ Income}$	Min
	C9	C-D	$\frac{Total\ Advances}{Total\ Deposits}$	Min
Profitability	C10	ROA	$\frac{Net\ Profit}{Total\ Assets}$	Max
	C11	ROE	$\frac{Net\ Profit}{Shareholders'\ Equity}$	Max

Table 2. Description of Selected Variables (Criteria). (Source: Authors' Compilation)

strength, asset quality, operational efficiency, and profitability. Therefore, the criteria or resilience indicators are grouped into four major dimensions: Solvency, Asset Quality, Efficiency, and Profitability, and treated as indicators of banks' financial resilience. Eleven financial ratios are incorporated under the four dimensions of the resilience framework. Table 2 presents the descriptions of the resilience indicators (variables). The variables are selected and established based on the available literature and related concepts and theories on banks' performance and financial resilience[3, 4, 37, 38, 43, 55–63].

3.1. MEREC Approach

Keshavarz et al.[22] proposed a new objective weighting method, MEREC, which utilizes the removal effect on alternatives to determine attribute weights. In contrast to previous criteria weight calculation techniques, the MEREC method relies on how the removal of conditions affects the total effects of substitutes[64, 65]. Higher weights are assigned to attributes with greater performance effects, and smaller weights are given to attributes with smaller performance effects. It gives an innovative approach to the criteria weight calculation technique.

The method follows the following six steps[22, 64, 65]:

Step 1: Construct the decision/evaluation matrix. The multi-criteria decision-making incorporates an $m \times n$ matrix where m is the number of alternatives and n is the number of criteria

$$X = x_{ij} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \quad (1)$$

The elements of the matrix are denoted by x_{ij}, where $i = 1, \ldots, m$, and $j = 1, \ldots, n$. The values in the matrix should be greater than zero ($x_{ij} > 0$) as well as a positive integer. Subsequently, any negative criteria values will be adjusted using the proper method.

Step 2: Normalize the decision matrix n_{ij}

Simple linear normalization procedure is applied to scale the elements of the decision matrix.

$$n_{ij} = \frac{x_{ij}}{\max x_j} \text{ if } j \in B \tag{2}$$

$$n_{ij} = \frac{\min x_j}{x_{ij}} \text{ if } j \in C \tag{3}$$

Where B is the benefit group of criteria (Max-type), and C represents the cost criteria (Min-type)

Step 3: Calculate the overall performance P_i

In this step, a logarithmic aggregation measure with equal weights for all criteria is employed to compute the overall performance values P_i. This measure is based on a non-linear function. Based on the normalized decision matrix obtained in the previous step, larger values of n_{ij} indicate better performance of the alternatives.

$$P_i = \ln\left(1 + \left(\frac{1}{n}\sum_{j=1}^{n} |\ln(n_{ij})|\right)\right) \tag{4}$$

Step 4: Determine the performance of alternatives by removing each criterion

This step evaluates the effect of each criterion on the overall performance of the alternatives by excluding one criterion at a time. Let P'_{ij} denote the overall performance of the i^{th} alternative when the j^{th} criterion is removed.

The performance value after removing the j^{th} criterion is computed as:

$$P'_{ij} = \ln\left(1 + \left(\frac{1}{n-1}\sum_{k=1,\, k\neq j}^{n} |\ln(n_{ik})|\right)\right) \tag{5}$$

Step 5: The summation of absolute deviations

In this step, we calculate the removal effect of the j^{th} criterion based on the values obtained in steps 3 and 4. Let R_j is the removal effect of the j^{th} criterion

$$R_j = \sum_{i=1}^{m} |P'_{ij} - P_i| \tag{6}$$

Step 6: Determine the final weights of the criterion

$$W_j = \frac{R_j}{\sum_{j=1}^{n} R_j} \tag{7}$$

3.2. RAM Method

The Root Assessment Method (RAM) aims to derive the utility value of each alternative by aggregating its scores over decision criteria. The Method was first applied to sustainability-focused multi-criteria decision problems, addressing complex evaluations with transparent ranking logic[20]. Subsequent studies extended its application to energy technology selection, manufacturing systems, urban and cost-of-living assessment, and comparative methodological analysis, including fuzzy, spherical fuzzy, and neutrosophic environments to manage uncertainty[21, 66–68]. The RAM method is selected in this bank resilience study, considering its low computational complexity, stable rankings[20], and effective integration with MEREC-based objective weighting, which may outperform conventional outranking MCDM techniques in complex financial evaluations.

The method involves the following major steps[20, 21, 68]:

Step 1: The first step involves forming an initial matrix X. This is preceded by defining the m set of alternatives A_i and a set of n criteria C_i

$$D = x_{ij} = \begin{array}{c} \\ A_1 \\ A_2 \\ \vdots \\ \vdots \\ A_m \end{array} \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{m\times n} \end{bmatrix} \tag{8}$$

Where alternatives i = 1, 2, . . . , m and criterion j = 1, 2, . . . ,n canbe expressed based on the nature of the criteria, i.e., cost (C) or benefit (B).

Step 2: Normalization of the initial matrix (x) using the linear sum normalization

$$n_{ij} = \frac{x_{ij}}{\sum_{i=1}^{m} x_{ij}} \tag{9}$$

Where elements x_{ij} represents the elements of the matrix x.

Step 3: Determination of the weighted normalized matrix V

$$V_{ij} = n_{ij} \times w_j \tag{10}$$

Where n_{ij} represents the normalized values calculated in step 2, and w_j represents the MEREC weights.

Step 4: Calculate the aggregates of weighted normalized values of beneficial and non-beneficial criteria for each alternative using the following equations

$$K_{+i} = \sum_{j=1}^{n} v_{+ij} \tag{11}$$

$$K_{-i} = \sum_{j=1}^{n} v_{-ij} \tag{12}$$

Whereas K_{+i} is the sum of weighted normalized values of the beneficial criteria and K_{-i} is the sum of weighted normalized values of the cost criteria.

Step 5: Determination of the aggregate relative resilience scores of alternatives using the following function Q_i

$$Q_i = \sqrt[2+K]{2 + K_{+i}} \tag{13}$$

The ranking of alternatives relies on the final values Q_i in descending order.

3.3. Mann–Kendall Trend Analysis

To determine whether the financial resilience of individual banks exhibited a monotonic trend over the study period, the univariate Mann–Kendall non-parametric trend test was applied to the composite RAM-based financial resilience scores for each bank from 2013 to 2024. The Mann–Kendall test is widely used for trend detection in time-series data because it does not require assumptions of normality and is robust to outliers.

The test statistic is computed as:

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \text{sgn}(x_j - x_i) \tag{14}$$

Where n is the number of years, x_i, and x_j denote the composite RAM scores observed at time points i and j, respectively, and the sign function sgn(·) is defined as:

$$\text{sgn}(x_j - x_i) = \begin{cases} +1, & \text{if } x_j - x_i > 0, \\ 0, & \text{if } x_j - x_i = 0, \\ -1, & \text{if } x_j - x_i < 0. \end{cases}$$

For sample sizes greater than ten, the statistic S is standardized to a normal variate Z,

$$\text{Where, } Z = \begin{cases} \frac{S-1}{\sqrt{\text{Var}(S)}} & \text{if } S > 0 \\ 0 & \text{if } S = 0 \\ \frac{S+1}{\sqrt{\text{Var}(S)}} & \text{if } S < 0 \end{cases}$$

Here,

$$\text{Var}(S) = \frac{n(n-1)(2n+5)}{18}$$

It tests the following hypotheses:

H_0: No monotonic trend exists in the series
H_1: A monotonic trend exists in the series

A positive and statistically significant Z-value indicates an increasing trend in financial resilience, whereas a negative and statistically significant Z-value indicates a decreasing trend. Statistical significance is evaluated at the 5% level ($p < 0.05$).

4. Results and Discussions

4.1. Analysis using MEREC

The study applies a multi-stage MCDM-integrated Mann-Kendall model. Initially, a 29 × 11 ($m \times n$) decision matrix is developed based on 11 criteria related to 29 banks (alternatives), for each year from 2013-2014 to 2024-2025, using Eq. (1). The relative importance of each criterion or financial resilience indicator is determined using the MEREC weighting method. Then, the weights obtained are integrated into the RAM approach to generate the resilience index. The ranking of banks is determined based on the aggregate function calculated at the end of the RAM approach.

The values in the decision matrix (Eq. (1)) were normalized for each period using Eq. (2) using the linear min-max normalization method to convert all the cost-benefit criteria into the same measurement scale. Subsequently, a logarithmic aggregation measure based on equal criterion weights is applied to the normalized values n_{ij} to compute the P_i values in Eq. (4). The p_i values represent the initial overall performance scores of the alternatives, taking into account all criteria. Accordingly, the modified performance values P'_{ij} are calculated using Eq. (5) by removing one criterion at a time from the set of criteria associated with alternatives, i.e., column-wise elimination. Equation (6) is then used to evaluate the criterion removal effect R_j by measuring the absolute deviation between P'_{ij} and P_i for each alternative i and criterion j. Finally, the criterion weights W_j for each study period are determined using Eq. (7) and are presented in Table 3.

The MEREC results, as shown in Table 3, reveals moderate variation in criteria weights over time. CAR (C1) increases from 0.046 (2013) to 0.123 (2022), reflecting rising regulatory emphasis. Loan growth (C6) fluctuates between 0.064 and 0.092, indicating cyclical sensitivity. Cost efficiency (C8) also varies (0.069–0.098), while ROA (C10) and ROE (C11) remain relatively stable (0.088–0.102). This suggests limited differentiation and a consistently lower role in resilience assessment.

Figure 2. Relative Importance of Financial Resilience Indicators (2013-2024). (Source: Estimated by the authors)

Fig 2 highlights temporal shifts in criteria importance as well as fluctuations in loan growth and efficiency indicators. This reflects impact of real-world events such as the AQR and IBC phases. On the other hand relative weight stability in profitability indicators suggests consistent regulatory and operational frameworks.

4.2. Analysis using RAM

The method RAM evaluated banking resilience by aggregating multiple financial indicators in a structured manner. The first step is to build the initial decision matrix D (Eq. (8)), which summarizes the performance of each banking alternative over all selected resilience indicators. Next it applies linear sum normalization (Eq. (9)) to enable logical comparison across indicators measured in different units. The normalized matrix is then weighted (Eq. (10)) with the MEREC-derived weight coefficients shown in Table 3 to produce the weighted normalized matrix V, which ensures that criteria with higher systemic relevance have a stronger effect on the evaluation.

RAM explicitly accounts for the asymmetric roles of beneficial and non-beneficial criteria by aggregating them separately, as expressed in Equations(11) and (12). K_{+i} is the sum of weighted normalized values of the beneficial criteria and K_{-i} is the sum of weighted normalized values of the cost criteria which are presented in Table 4. Beneficial criteria contribute positively to resilience, whereas non-beneficial criteria capture vulnerability channels. This separation enables the final RAM utility score Q_i, computed using Eq. (13), to reflect the balance between shock-absorption capacity and risk exposure.

Table 5 shows the RAM-based aggregate relative resilience scores (Q_i) of each bank from 2013 to 2024. The (Q_i) scores show a high level of consistency. The values generally fall between about 1.41 and 1.42.

The Q_i values are normalized and presented in Table 6 which shows the financial resilience ranges between 0.995 and

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11
2013	0.046	0.101	0.095	0.095	0.091	0.082	0.101	0.082	0.103	0.102	0.102
2014	0.060	0.100	0.083	0.093	0.090	0.079	0.097	0.098	0.101	0.100	0.100
2015	0.043	0.103	0.094	0.098	0.096	0.091	0.103	0.069	0.103	0.100	0.101
2016	0.073	0.099	0.084	0.091	0.089	0.077	0.099	0.096	0.098	0.097	0.097
2017	0.047	0.101	0.093	0.097	0.095	0.084	0.102	0.078	0.102	0.100	0.100
2018	0.041	0.103	0.096	0.098	0.097	0.092	0.087	0.080	0.102	0.102	0.102
2019	0.054	0.100	0.093	0.096	0.094	0.087	0.102	0.074	0.101	0.101	0.101
2020	0.090	0.096	0.092	0.089	0.087	0.064	0.098	0.093	0.097	0.096	0.097
2021	0.065	0.099	0.095	0.094	0.093	0.064	0.101	0.097	0.100	0.096	0.096
2022	0.123	0.091	0.078	0.086	0.085	0.084	0.095	0.088	0.093	0.088	0.090
2023	0.048	0.101	0.098	0.100	0.098	0.077	0.103	0.073	0.103	0.100	0.099
2024	0.100	0.093	0.080	0.090	0.090	0.080	0.097	0.091	0.096	0.092	0.091

Table 3. Year-Wise Obtained Criteria Weights using the MEREC Approach. (Source: Estimated by the authors)

	2013		**2014**		**2015**		**2016**		**2017**		**2018**		**2019**		**2020**		**2021**		**2022**		**2023**		**2024**	
	K_{+i}	K_{-i}	K_{+i}	K_{-i}	K_{+i}	K_{-i}	K_{+i}	K_{-i}	K_{+i}	K_{-i}	K_{+i}	K_{-i}	K_{+i}	K_{-i}	K_{+i}	K_{-i}	K_{+i}	K_{-i}	K_{+i}	K_{-i}	K_{+i}	K_{-i}	K_{+i}	K_{-i}
CB1	0.019	0.015	0.019	0.015	0.021	0.018	0.018	0.018	0.017	0.022	0.014	0.020	0.017	0.019	0.018	0.016	0.019	0.017	0.014	0.014	0.018	0.018	0.018	0.013
CB2	0.015	0.018	0.015	0.014	0.011	0.019	0.015	0.018	0.015	0.021	0.013	0.019	0.015	0.024	0.015	0.018	0.014	0.020	0.017	0.014	0.015	0.019	0.016	0.018
CB3	0.013	0.023	0.013	0.020	0.009	0.026	0.013	0.025	0.012	0.024	0.012	0.023	0.015	0.023	0.016	0.018	0.013	0.020	0.014	0.019	0.013	0.025	0.014	0.023
CB4	0.016	0.021	0.019	0.023	0.012	0.023	0.009	0.027	0.013	0.021	0.005	0.025	0.016	0.019	0.016	0.014	0.015	0.014	0.019	0.010	0.020	0.013	0.021	0.015
CB5	0.012	0.020	0.010	0.063	0.008	0.010	0.015	0.018	0.013	0.012	0.012	0.011	0.020	0.015	0.020	0.015	0.027	0.014	0.026	0.008	0.021	0.015	0.019	0.022
CB6	0.014	0.020	0.014	0.017	0.011	0.021	0.015	0.022	0.013	0.022	0.011	0.021	0.015	0.020	0.015	0.023	0.014	0.020	0.017	0.018	0.016	0.022	0.017	0.020
CB7	0.010	0.027	0.013	0.022	0.009	0.023	0.010	0.024	0.008	0.024	0.007	0.022	0.014	0.025	0.015	0.020	0.012	0.024	0.012	0.017	0.013	0.021	0.015	0.018
CB8	0.018	0.012	0.019	0.011	0.021	0.013	0.021	0.012	0.021	0.013	0.016	0.014	0.017	0.014	0.019	0.013	0.020	0.019	0.020	0.017	0.018	0.024	0.019	0.020
CB9	0.018	0.016	0.019	0.015	0.021	0.014	0.019	0.014	0.020	0.015	0.016	0.014	0.018	0.015	0.018	0.017	0.016	0.019	0.016	0.015	0.014	0.023	0.014	0.027
CB10	0.006	0.022	0.008	0.032	0.011	0.011	0.016	0.026	0.016	0.013	0.013	0.016	0.017	0.014	0.015	0.030	0.009	0.021	0.011	0.022	0.008	0.017	0.009	0.021
CB11	0.019	0.012	0.019	0.012	0.018	0.014	0.018	0.014	0.019	0.016	0.015	0.015	0.017	0.014	0.017	0.012	0.017	0.015	0.017	0.012	0.016	0.017	0.016	0.015
CB12	0.019	0.014	0.020	0.011	0.022	0.015	0.021	0.012	0.021	0.014	0.017	0.015	0.021	0.014	0.020	0.011	0.023	0.012	0.021	0.009	0.018	0.020	0.018	0.014
CB13	0.019	0.020	0.019	0.020	0.020	0.022	0.019	0.020	0.019	0.021	0.015	0.019	0.017	0.016	0.019	0.012	0.023	0.013	0.022	0.010	0.020	0.016	0.020	0.013
CB14	0.015	0.018	0.015	0.014	0.015	0.016	0.017	0.016	0.017	0.019	0.013	0.018	0.016	0.018	0.017	0.025	0.014	0.017	0.015	0.013	0.014	0.017	0.017	0.015
CB15	0.012	0.024	0.011	0.024	0.008	0.034	0.009	0.032	0.007	0.028	0.011	0.030	0.010	0.020	0.017	0.016	0.015	0.018	0.014	0.019	0.014	0.020	0.016	0.018
CB16	0.019	0.016	0.018	0.016	0.022	0.017	0.021	0.014	0.021	0.017	0.015	0.020	0.017	0.017	0.018	0.013	0.019	0.014	0.020	0.011	0.018	0.017	0.011	0.017
CB17	0.019	0.012	0.015	0.014	0.017	0.018	0.009	0.017	0.016	0.016	0.013	0.019	0.013	0.017	0.015	0.016	0.013	0.022	0.016	0.016	0.015	0.017	0.017	0.017
CB18	0.015	0.017	0.016	0.014	0.017	0.013	0.018	0.014	0.017	0.015	0.014	0.016	0.016	0.016	0.016	0.012	0.014	0.016	0.018	0.013	0.018	0.021	0.016	0.019
CB19	0.017	0.014	0.018	0.011	0.020	0.012	0.018	0.012	0.018	0.015	0.015	0.018	0.017	0.016	0.018	0.016	0.017	0.016	0.019	0.009	0.018	0.013	0.019	0.016
CB20	0.020	0.015	0.020	0.015	0.021	0.022	0.021	0.013	0.022	0.014	0.018	0.015	0.020	0.013	0.021	0.010	0.024	0.012	0.024	0.008	0.021	0.013	0.022	0.012
CB21	0.012	0.022	0.013	0.020	0.015	0.018	0.014	0.020	0.013	0.020	0.076	0.022	0.012	0.025	0.010	0.029	0.014	0.022	0.015	0.022	0.011	0.023	0.014	0.025
CB22	0.014	0.023	0.014	0.021	0.010	0.028	0.014	0.021	0.008	0.026	0.008	0.023	0.016	0.022	0.015	0.024	0.010	0.024	0.011	0.021	0.012	0.020	0.018	0.019
CB23	0.016	0.032	0.017	0.022	0.018	0.024	0.018	0.018	0.019	0.017	0.015	0.019	0.017	0.018	0.017	0.014	0.010	0.016	0.014	0.015	0.013	0.020	0.012	0.015
CB24	0.016	0.014	0.014	0.011	0.015	0.016	0.016	0.013	0.016	0.015	0.013	0.018	0.014	0.019	0.014	0.018	0.009	0.026	0.013	0.025	0.015	0.022	0.016	0.022
CB25	0.015	0.023	0.015	0.017	0.016	0.023	0.016	0.020	0.015	0.023	0.012	0.021	0.016	0.018	0.016	0.018	0.015	0.019	0.016	0.015	0.015	0.021	0.015	0.017
CB26	0.017	0.015	0.018	0.010	0.020	0.012	0.019	0.013	0.019	0.011	0.016	0.015	0.019	0.012	0.020	0.014	0.024	0.014	0.025	0.008	0.022	0.015	0.024	0.021
CB27	0.015	0.024	0.014	0.019	0.008	0.028	0.011	0.024	0.009	0.028	0.008	0.026	0.013	0.022	0.015	0.018	0.010	0.017	0.012	0.092	0.012	0.021	0.012	0.022
CB28	0.014	0.022	0.014	0.021	0.014	0.022	0.014	0.025	0.011	0.024	0.010	0.025	0.014	0.025	0.015	0.029	0.014	0.023	0.015	0.018	0.016	0.019	0.018	0.016
CB29	0.019	0.017	0.019	0.017	0.021	0.019	0.020	0.012	0.019	0.025	0.013	0.026	0.006	0.035	0.012	0.030	0.012	0.039	0.010	0.026	0.008	0.019	0.011	0.019

Table 4. Aggregates of Weighted Normalized Values of the Beneficial (K_{+i}) and Cost (K_{-i}) Criteria

1. This normalization simplified it to compare the scores across banks and over the years. Most banks keep consistently high normalized scores, reflecting strong and stable performance throughout the study period. The limited spread in normalized Q_i values highlights the stability of the banking system, while the observed differences give valuable insights into how banks perform relative to each other. This supports the reliability of the RAM-based evaluation and suggests that the banking system exhibits consistent resilience with limited dispersion in performance levels.

The banks such as CB8, CB11, CB12, and CB20 consistently reflected higher Q_i scores across multiple years. These banks demonstrate minimal fluctuations, reflecting strong adaptability and sustained resilience under uncertainties. In contrast, banks such as CB5, CB6, CB10, and CB15 demonstrates slightly lower Q_i values and maximum variation over time. Although their scores remain within a high range, the observed fluctuations suggest relatively lower stability and sensitivity to macro-level changes compared to the top-performing banks.

Accordingly, Table 7 presents the year-wise ranking results obtained using the MEREC-RAM approach.

The annual RAM-based rankings provide additional insight about banks' relative financial resilience over time. While the changes in normalized Q_i values between banks are almost negligible, the modest variations discovered have a substantial impact on year-over-year rankings. The narrow dispersion of RAM scores reflects the standardized and highly regulated nature of the Indian banking system under RBI and Basel norms. This demonstrates that even slight changes in performance can affect how banks compare in a competitive market.

A set of banks frequently holds the top ranks. In For instance, Kotak Mahindra Bank Ltd. (CB20) performed exceptionally well, frequently ranking among the top places and achieving first position multiple times, particularly in recent years. Similarly, Tamilnad Mercantile Bank Ltd. (CB26), HDFC Bank Ltd. (CB12) and ICICI Bank Ltd. (CB13) maintained stable ranks throughout the study period. City Union Bank Ltd. (CB8) and Federal Bank Ltd. (CB11) also exhibited relatively high and stable rankings, with minor fluctuations. These banks can be considered as consistently resilient performers within the sample. On the other hand, CSB Bank Ltd. (CB5) reflected significant improvement in rankings from bottom to top ten over the years. Bank of Maharashtra (CB4) also demonstrated significant improvement in rankings by reaching to the second place during 2024.

	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024
CB1	1.417	1.417	1.417	1.416	1.415	1.414	1.416	1.417	1.417	1.416	1.416	1.417
CB2	1.415	1.416	1.414	1.415	1.414	1.414	1.414	1.415	1.414	1.416	1.415	1.415
CB3	1.413	1.414	1.411	1.413	1.413	1.413	1.414	1.415	1.414	1.414	1.413	1.414
CB4	1.415	1.415	1.413	1.411	1.413	1.410	1.415	1.416	1.416	1.419	1.418	1.418
CB5	1.414	1.403	1.414	1.415	1.416	1.416	1.418	1.418	1.420	1.421	1.418	1.416
CB6	1.414	1.415	1.413	1.414	1.413	1.413	1.414	1.414	1.414	1.416	1.415	1.415
CB7	1.411	1.413	1.412	1.412	1.411	1.411	1.413	1.414	1.413	1.414	1.413	1.415
CB8	1.418	1.418	1.418	1.419	1.418	1.417	1.417	1.418	1.417	1.417	1.415	1.416
CB9	1.417	1.417	1.418	1.417	1.418	1.416	1.417	1.416	1.415	1.416	1.413	1.413
CB10	1.411	1.409	1.415	1.413	1.417	1.415	1.417	1.412	1.412	1.413	1.413	1.412
CB11	1.418	1.418	1.417	1.417	1.417	1.416	1.417	1.417	1.417	1.417	1.416	1.416
CB12	1.417	1.418	1.418	1.419	1.418	1.417	1.418	1.419	1.419	1.420	1.416	1.417
CB13	1.416	1.416	1.416	1.416	1.416	1.415	1.416	1.418	1.419	1.420	1.417	1.418
CB14	1.415	1.416	1.416	1.416	1.416	1.414	1.415	1.414	1.415	1.416	1.415	1.417
CB15	1.413	1.412	1.409	1.409	1.410	1.411	1.413	1.416	1.415	1.415	1.414	1.415
CB16	1.417	1.417	1.418	1.418	1.417	1.415	1.416	1.417	1.417	1.418	1.416	1.414
CB17	1.418	1.416	1.416	1.413	1.416	1.414	1.415	1.416	1.413	1.416	1.415	1.416
CB18	1.415	1.416	1.417	1.417	1.416	1.415	1.416	1.417	1.415	1.417	1.415	1.415
CB19	1.417	1.418	1.418	1.418	1.417	1.415	1.416	1.416	1.417	1.419	1.417	1.417
CB20	1.417	1.418	1.416	1.418	1.419	1.417	1.418	1.419	1.420	1.421	1.418	1.419
CB21	1.413	1.414	1.415	1.414	1.414	1.435	1.413	1.411	1.414	1.414	1.412	1.413
CB22	1.414	1.414	1.411	1.414	1.411	1.411	1.414	1.414	1.412	1.413	1.413	1.416
CB23	1.412	1.415	1.415	1.416	1.417	1.415	1.416	1.417	1.414	1.415	1.414	1.415
CB24	1.416	1.416	1.416	1.417	1.416	1.414	1.415	1.415	1.411	1.413	1.414	1.414
CB25	1.414	1.415	1.414	1.415	1.414	1.413	1.415	1.415	1.415	1.416	1.414	1.415
CB26	1.417	1.418	1.418	1.418	1.418	1.416	1.418	1.418	1.419	1.421	1.418	1.417
CB27	1.414	1.414	1.410	1.412	1.411	1.411	1.414	1.415	1.413	1.397	1.413	1.413
CB28	1.414	1.414	1.414	1.413	1.412	1.412	1.413	1.412	1.414	1.415	1.415	1.417
CB29	1.417	1.417	1.417	1.418	1.415	1.412	1.408	1.411	1.409	1.412	1.412	1.413

Table 5. Aggregate resilience scores Q_i of banks for each year from 2013-2024. (Source: Authors' Computation)

In contrast, several banks show persistent lower rankings, indicating relatively weaker resilience. For instance, Bank of India (CB3), Central Bank of India (CB7) and Dhanlaxmi Bank Ltd. (CB10), UCO Bank (CB27) frequently appeared in the lower ranks. This indicates continuous structural or performance challenges. These banks also exhibit higher ranking divergence reflecting instability in their resilience over time. On the other hand, Yes Bank Ltd. (CB29) experienced substantial declines in their rankings especially from 2018, indicative of the existence of underlying challenges and different government's correction measures such as AQR and IBC implementations during 2016-2017, as well as interventions with revised Prompt Corrective Action (PCA).

Figure 3 presents the heatmap on normalized resilience score of Qi across different period. The colour gradient, ranging from darker shades (lower scores) to lighter (higher scores) provides a visual representation of the variation in financial resilience across banks and over time. The colour gradient, ranging from darker shades (lower scores) to lighter shades (higher scores), provides a visual representation of the variation in financial resilience across banks and over time. A noticeable deviation is observed around 2018, where a darker vertical band appears across a large number of banks. This indicates a sector-wide decline in resilience during that year, pointing to the presence of a systemic shock affecting the entire banking system rather than isolated bank-specific issues. This might be manifested from two major events in the Indian financial system e.g., PNB scam and the IL&FS crisis during the year 2018 which significantly affected the entire system.

Figure 3. Heatmap on banks year-wise composite financial resilience scores obtained using MEREC-RAM approach. (Source: Authors' Computation)

Following this period, it shows a gradual return to normal range, reflecting recovery and improved resilience in subsequent

	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024
CB1	0.999	0.999	0.999	0.998	0.997	0.985	0.998	0.998	0.997	0.996	0.998	0.999
CB2	0.998	0.998	0.997	0.998	0.997	0.985	0.997	0.997	0.996	0.997	0.998	0.997
CB3	0.997	0.997	0.995	0.996	0.996	0.984	0.997	0.997	0.995	0.995	0.996	0.996
CB4	0.997	0.998	0.996	0.995	0.996	0.983	0.998	0.998	0.997	0.998	1.000	0.999
CB5	0.997	0.989	0.997	0.997	0.998	0.986	0.999	0.999	1.000	1.000	1.000	0.998
CB6	0.997	0.998	0.996	0.997	0.996	0.985	0.997	0.996	0.996	0.996	0.997	0.997
CB7	0.995	0.996	0.995	0.995	0.995	0.983	0.996	0.997	0.995	0.995	0.997	0.997
CB8	1.000	1.000	1.000	1.000	1.000	0.987	0.999	0.999	0.997	0.997	0.997	0.998
CB9	0.999	0.999	1.000	0.999	0.999	0.987	0.999	0.998	0.996	0.996	0.996	0.996
CB10	0.995	0.994	0.998	0.996	0.999	0.986	0.999	0.995	0.994	0.994	0.996	0.995
CB11	1.000	1.000	0.999	0.999	0.999	0.987	0.999	0.999	0.997	0.997	0.998	0.998
CB12	0.999	1.000	1.000	1.000	1.000	0.987	1.000	0.999	0.999	0.999	0.998	0.999
CB13	0.998	0.998	0.998	0.998	0.998	0.986	0.999	0.999	0.999	0.999	0.999	0.999
CB14	0.998	0.998	0.998	0.998	0.998	0.985	0.998	0.996	0.996	0.996	0.998	0.998
CB15	0.996	0.996	0.993	0.994	0.994	0.983	0.996	0.998	0.996	0.995	0.997	0.998
CB16	0.999	0.999	1.000	1.000	0.999	0.986	0.998	0.999	0.998	0.998	0.999	0.996
CB17	1.000	0.998	0.998	0.996	0.998	0.985	0.997	0.997	0.995	0.996	0.998	0.998
CB18	0.998	0.999	0.999	0.999	0.999	0.986	0.998	0.998	0.996	0.997	0.998	0.997
CB19	0.999	0.999	1.000	0.999	0.999	0.986	0.999	0.998	0.997	0.998	0.999	0.999
CB20	0.999	0.999	0.998	1.000	1.000	0.987	1.000	1.000	1.000	0.999	1.000	1.000
CB21	0.996	0.997	0.998	0.997	0.997	1.000	0.996	0.994	0.995	0.995	0.996	0.996
CB22	0.997	0.997	0.995	0.997	0.994	0.983	0.997	0.996	0.994	0.994	0.997	0.998
CB23	0.996	0.997	0.997	0.998	0.999	0.986	0.998	0.998	0.996	0.996	0.997	0.997
CB24	0.999	0.999	0.998	0.999	0.998	0.985	0.997	0.997	0.993	0.994	0.997	0.997
CB25	0.997	0.998	0.997	0.998	0.997	0.985	0.998	0.997	0.996	0.996	0.997	0.997
CB26	0.999	1.000	1.000	0.999	1.000	0.987	1.000	0.999	0.999	1.000	1.000	0.999
CB27	0.997	0.997	0.994	0.996	0.995	0.983	0.997	0.997	0.995	0.983	0.996	0.996
CB28	0.997	0.997	0.997	0.996	0.995	0.984	0.996	0.995	0.995	0.996	0.998	0.998
CB29	0.999	0.999	0.999	1.000	0.997	0.984	0.993	0.994	0.992	0.993	0.996	0.996

Table 6. Normalized values of aggregate resilience scores (Q_i). (Source: Estimated by the authors)

years. This pattern highlights the ability of banks to adapt and regain stability after adverse conditions.

After this period, it indicates a gradual recovery back to normal range, which shows that banks have returned to a more stable situation in later years. Interestingly, while most other banks' resilience scores fell significantly, as seen in Fig. 3, the Punjab & Sind Bank (CB21) had the highest ranking in terms of resilience in 2018 (Table 7), which was likely due to a one-off balance-sheet adjustments. Overall, the RAM rankings highlight meaningful improvements in relative resilience and systemic stress, rather than dramatic shifts in the banks' absolute strength.

4.3. Hypotheses Testing

The Mann–Kendall (MK) non-parametric trend test was employed using Eq. (14) to examine both bank-level and sector-level dynamics in composite RAM-based financial resilience scores over the period 2013–2024. First, annual RAM scores for 29 banks were structured as a balanced time series, enabling univariate MK tests for each bank to detect statistically significant monotonic trends. The test statistic S, Kendall's tau (τ), and associated p-values were computed for all banks which allow for classification of resilience trajectories as increasing or decreasing trend.

The Mann–Kendall trend analysis in Fig. 4 demonstrates the upward (increasing) and downward (decreasing) trend in financial resilience, as well as banks with no significant trend or (stable over the period) highlighted from 2013 to 2024. The

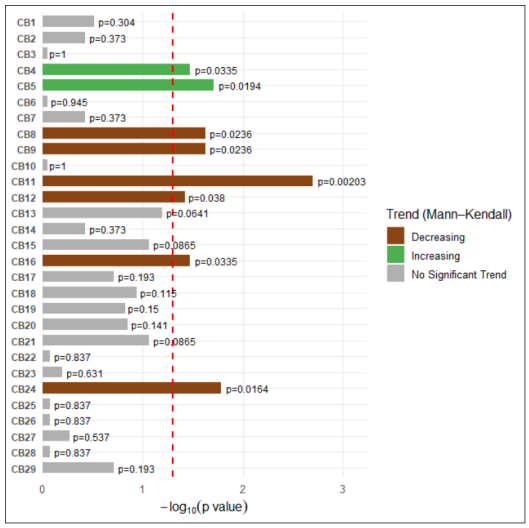


Figure 4. Bank-specific Mann–Kendall Trend Test results (2013–2024). (Source: Authors' Computation)

MEREC–RAM rankings presented in Table 6 capture the relative resilience levels of banks in a particular year, while the Mann–Kendall test captures the dynamics of resilience.

The results reveal that the majority of banks do not exhibit statistically significant trends in their resilience scores

	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024
CB1	5	7	7	14	18	16	14	10	8	17	8	5
CB2	15	16	22	16	19	19	22	18	17	12	15	16
CB3	24	24	26	25	24	22	23	17	20	23	27	24
CB4	17	17	23	28	22	29	17	13	11	6	3	2
CB5	22	29	19	18	13	8	4	5	1	1	4	14
CB6	18	19	24	19	23	21	20	24	18	18	17	20
CB7	28	26	25	27	26	26	25	22	25	22	22	18
CB8	3	2	1	2	2	4	5	6	7	11	16	11
CB9	9	8	5	8	5	5	7	14	14	14	24	28
CB10	29	28	17	22	10	11	6	27	26	27	26	29
CB11	2	3	8	9	7	7	8	7	10	10	9	12
CB12	6	1	2	1	4	3	1	2	3	4	10	6
CB13	13	15	12	15	16	14	9	3	5	5	5	3
CB14	16	14	14	13	15	15	15	23	15	15	14	9
CB15	26	27	29	29	29	27	27	15	13	21	18	15
CB16	7	9	6	5	6	13	11	8	6	8	7	23
CB17	1	13	15	23	14	18	18	16	24	16	12	13
CB18	14	12	9	10	11	9	13	9	12	9	11	19
CB19	8	5	3	7	8	10	10	12	9	7	6	7
CB20	4	6	11	3	1	2	2	1	2	3	1	1
CB21	25	25	16	20	20	1	28	29	21	24	28	27
CB22	23	22	27	21	28	25	21	25	27	26	23	10
CB23	27	20	18	12	9	12	12	11	19	19	21	21
CB24	12	11	13	11	12	17	19	21	28	25	20	22
CB25	19	18	20	17	21	20	16	19	16	13	19	17
CB26	10	4	4	6	3	6	3	4	4	2	2	4
CB27	20	21	28	26	27	28	24	20	23	29	25	26
CB28	21	23	21	24	25	24	26	26	22	20	13	8
CB29	11	10	10	4	17	23	29	28	29	28	29	25

Table 7. Year-wise MEREC-based RAM rankings. (Source: Estimated by the authors)

over time. This suggests that financial resilience remains relatively stable for most institutions with no consistent upward or downward trajectory.

However, a subset of banks demonstrates statistically significant trends. Specifically, banks such as CB4 and CB5 exhibit significant improving trends, indicating gradual improvements in resilience over the study period. These results suggest that these banks have strengthened their financial position over time. In contrast, several banks display significant decreasing trends in resilience. Notably, CB8, CB9, CB11, CB12, CB16, and CB24 show statistically significant downward trends, with CB11 exhibiting the strongest decline ($p < 0.01$). This indicates a functional decline in financial resilience for these institutions, which may warrant closer attention from the side of regulatory bodies.

Banks CB4 and CB5 indicate statistically significant upward trends in resilience which align with Financial Resilience Theory and show signs of improvement in solvency, asset quality, efficiency, and profitability. Capital Buffer Theory focuses on the need to maintain sufficient capital and Tier-1 buffers to reduce bank fragility and promote long-term stability. This reflected in the banks with positive trends as well as with the banks showing stable financial resilience throughout the period. Meanwhile, banks that have experienced significant declines that correspond with the theory that weak buffers make banks more vulnerable due to problems with NPAs, inadequate provisioning, and earnings pressure. CB11 display marginally significant declining trend which reflect the lingering effects of capital erosion, governance challenges, and volatility in profitability. The majority of large public and private sector banks such as CB25, CB1, CB13, CB22, CB25, and so forth, showing no statistically significant trend implies relative stability over the period but limited monotonic improvement in resilience.

The sector-level resilience is further evaluated by constructing annual mean RAM scores, to which separate MK tests are applied. To assess heterogeneity in resilience patterns, the sample was stratified based on ownership structure (public vs. private banks). Both parametric and non-parametric tests were employed to determine whether resilience trajectories differed significantly across subgroups.

Although the initial bank-level Mann–Kendall analysis provides significant evidence on how some banks experienced increasing and decreasing trends in their financial resilience, the Mann–Kendall test for sector mean resilience indicates no statistically significant monotonic trend ($Z = -1.71, p = 0.086$). However, the negative tau (-0.394) suggests a weak declining tendency over the study period. Therefore, the null hypothesis (H_0^1) of no monotonic trend cannot be rejected at the aggregate level.

A Sen's slope estimator is used to quantify the magnitude and direction of sector-wide change. The estimated slope is visualised in Fig. 5. Initially, a Welch two-sample t-test is conducted on average RAM scores over the study period (2013–2024), which revealed a statistically significant difference between the two groups ($t = 4.64, p < 0.001$). A non-parametric Kruskal–Wallis test is additionally performed to ensure robustness. The results confirm a statistically significant difference

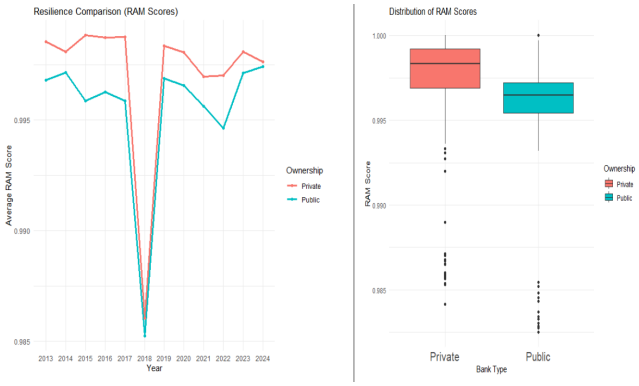


Figure 5. Ownership-Based Heterogeneity Analysis and Sectoral Trend in Financial Resilience. (Source: Authors' Computation)

in resilience scores, especially between the public and private sector banks ($\chi^2 = 71.096, p < 0.001$). The null hypothesis ($H_0^2$) of no difference between public and private banks is rejected, as Welch two-sample t-test and non-parametric tests Kruskal–Wallis reveal statistically significant differences in resilience across ownership groups.

As can be seen in the boxplot (Fig. 5), private banks have a more resilient recovery from the sharp drop around 2018, which is a sign of greater systemic resilience. The different recovery paths of public sector banks (PSBs) and private banks since 2018 show the impact of different governance structures, support for recapitalization, and operational flexibility. Public sector banks received significant recapitalization and governance changes through the EASE and Indradhanush programs. However, they also had to manage a larger amount of legacy non-performing assets. In contrast, private banks usually had more managerial freedom and better profit margin which allow them to recover and adapt more quickly. These differences in structure might caused the varying resilience of each sector after systemic shocks.

These validate that the observed variations in resilience are statistically significant and not an artifact of the MCDM-based ranking framework. The results are also highly consistent with Dynamic Capability Theory, which asserts that firms can maintain superior performance by sensing risks, seizing opportunities, and reconfiguring resources under changing environments[53]. The gradual recovery of the Indian banking sector after the structural change around 2018 also shows a positive adaptive response through a series of balance-sheet adjustments, capital augmentation, and operational restructuring.

4.4. Sensitivity Analysis and Robustness Test

To examine the effect of changes in criteria weighting, a sensitivity analysis is conducted by comparing the baseline MEREC–RAM results with alternative equal-weighting and Entropy-based weighting scenarios. The equal-weighting was implemented by applying 1/11, in which all financial resilience indicators were assigned identical importance 0.0909 (i.e., for 11 criteria), thereby eliminating any data-driven or subjective prioritisation among criteria. Correspondingly, the method entropy is applied to determine objectives which then followed by entropy-RAM rankings.

The study applied additional outranking techniques such as TOPSIS, RATMI, and MARCOS for a robustness test incorporating the MEREC weights. The Kendall's tau coefficient is employed to assess the degree of rank concordance among the ranking outcomes obtained from variations in weighting schemes and multiple outranking models. Fig. 6 presents the year-wise p-values of Kendall's tau coefficient for rank agreement in both the sensitivity analysis and robustness tests.

The consistently high Kendall's tau coefficients across all years (Fig. 5) indicate that the integrated approach produces stable and reliable rankings. Entropy and MEREC shows near-perfect agreement ($\tau = 0.96 - 1.00$), while Equal Weighting and MEREC ranges from 0.71 to 0.94. This demonstrates that the results are not driven by a specific weighting technique, but rather reflect inherent data structure. Under the robustness test, RAM and TOPSIS (0.79–0.95) show strongest agreement, followed by RAM-MARCOS (0.67–0.89) and RAM-RATMI (0.67–0.81). The absence of significant rank reversals across methods and time reinforces the robustness of the framework.

5. Conclusions

Assessing banks' financial resilience is vital in emerging economies, where financial systems are characterized by volatility, governance constraints, and heightened exposure to NPAs. Thus, in order to assess the financial resilience Indian commercial banks between 2013 and 2024, this study developed an integrated analytical framework that combines the MEREC objective weighting approach, the RAM-based composite resilience index, , and Mann-Kendall trend analysis.

The results showed that the RAM-based resilience rankings are narrowly clustered among banks, indicating the effect of structural convergence brought about by regulatory standardization. The statistical significance tests Kruskal–Wallis verify that the observed differences are meaningful and significant. However, the trend analysis shows that resilience trajectories at the bank level vary significantly. A limited number of banks have statistically significant monotonic trends, but cross-sectional rankings show year-to-year variability for most banks. The lack of a noticeable trend at the sector-level indicates that the Indian banking sector resilience develops through periodic adjustments rather than steady directional improvement. Moreover, while Private sector banks showed greater adaptive capacity and quicker recovery from systemic shocks, public banks were more sensitive to regulatory interventions.

The MEREC-RAM ranking result showed Kotak Mahindra Bank Ltd. and Tamilnad Mercantile Bank Ltd. consistently showed strong resilience across the study period. Meanwhile the CSB Bank Ltd. reflected significant improvement in rankings

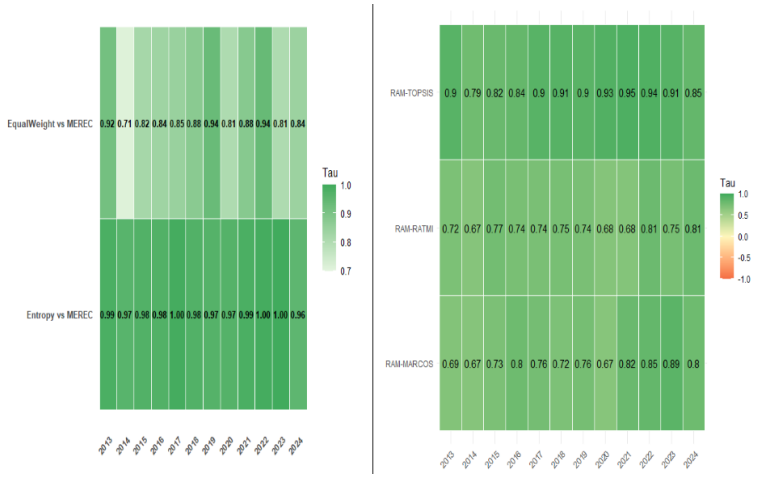


Figure 6. Heatmaps on Kendall's tau (τ) across methods. (Source: Authors' Computation)

from bottom to top ten over the years. Bank of Maharashtra also demonstrated significant improvement in rankings by reaching to the second place during 2024. Conversely, Bank of India, Central Bank of India and Dhanlaxmi Bank Ltd., UCO Bank constantly positioned at the bottom in the rankings. Interestingly, the Mann-kendall trend test showed significant improvement in the resilience of Bank of Maharashtra and CSB Bank Ltd. during the study period. Whereas, City Union Bank Ltd., DCB Bank Ltd., Federal Bank Ltd., HDFC Bank Ltd., Indusind Bank Ltd., and South Indian Bank Ltd. reflected significant decline in their financial resilience. However, other banks demonstrated stable or no significant improvement or deterioration.

This study contributes to Financial Resilience Theory in showing that resilience is non-linear, heterogeneous, and convergent rather than monotonic. These findings are also in line with Institutional Adaptation Theory, which contends that resilience outcomes depend on banks' capacity to adapt to shifting macroeconomic and regulatory conditions. The results are consistent with a convergence-differentiation approach to banking resilience, in which statistically significant performance differences resulting from differences in banks' capacity to adapt, manage risks, and react to changing conditions are contrasted with structural similarity brought about by regulatory frameworks.

The Mann-Kendall analysis shows that banks exhibit distinct resilience trajectories, which have significant managerial and policy significance. From a managerial perspective, the findings suggest that persistent adaptive capability is necessary in addition to maintaining high relative rankings. To achieve long-term resilience, banks must place the utmost importance on proactive risk management, ongoing balance-sheet reconfiguration, and operational effectiveness. Banks with increasing trends demonstrate great adaptive capacity and can serve as benchmarks for best practices, with a continuous emphasis on risk management and capital strengthening. Banks showing deteriorating trends (e.g., CB8, CB9, CB11, CB12, CB16, and CB24) should prioritize strengthening asset quality through stricter credit appraisal, early warning systems, and faster resolution of non-performing assets. Management should improve capital adequacy, liquidity buffers, and risk governance while adopting data-driven monitoring of profitability and operational efficiency. Diversification of loan portfolios and digital transformation can reduce concentration and operating risks. Regulators may encourage periodic stress testing, enhanced disclosure standards, and corrective action frameworks for vulnerable banks. From a policy perspective, the lack of a broad sector-wide trend indicates that while regulatory changes have stabilized the system, they have not consistently increased resilience.

Although the study makes important contributions, it is limited by its analysis to financial indicators without incorporating macroeconomic or market-based variables. Although the Mann–Kendall test captures monotonic trends but it unable to conduct a complete structural break or regime shifts. Further research could expand on this framework by including macro-financial variables, structural break or panel-based approaches, and identifying causal links between regulatory interventions, adaptive capacities, and banking system resilience.

Ethical Statement

No ethical approval was required for this study, as it did not involve human or animal subjects. The study relied exclusively on secondary data obtained from publicly available and authorized databases and annual reports.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Declaration of competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability Statements

The data supporting the findings of this study are available from the corresponding author upon reasonable request. However, the data are not publicly available due to access restrictions associated with the Prowess-IQ database and institutional data usage policies. Data were obtained from authorized sources

including Prowess-IQ (`https://prowess.cmie.com/`) and annual reports of banks.

Credit authorship contribution statement

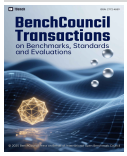
Priya Das: Conceptualization; Methodology; Investigation; Data Curation; Formal Analysis; Software; Visualization; Writing – Original Draft Preparation. Subir Kumar Sen: Conceptualization; Supervision; Project Administration; Validation; Visualization; Resources; Writing – Review & Editing.

References

1. F. Salignac, A. Marjolin, R. Reeve, and K. Muir, "Conceptualizing and measuring financial resilience," *Social Indicators Research*, vol. 145, no. 1, pp. 17–38, 2019, doi: https://doi.org/10.1007/s11205-019-02100-4.
2. F. Salignac, J. Hanoteau, and I. Ramia, "Financial resilience: a way forward towards economic development in developing countries," *Social Indicators Research*, vol. 160, no. 1, pp. 1–33, 2022, doi: https://doi.org/10.1007/s11205-021-02793-6.
3. S. K. Nkundabanyanga, E. Mugumya, I. Nalukenge, M. Muhwezi, and G. M. Najjemba, "Firm characteristics, innovation, financial resilience and survival of financial institutions," *Journal of Accounting in Emerging Economies*, vol. 10, no. 1, pp. 48–73, 2020, doi: https://doi.org/10.1108/JAEE-08-2018-0094.
4. F. S. Hamid, Y. J. Loke, and P. N. Chin, "Determinants of financial resilience: insights from an emerging economy," *Journal of social and economic development*, vol. 25, no. 2, pp. 479–499, 2023, doi: https://doi.org/10.1007/s40847-023-00239-y.
5. A. A. Tinta, I. M. Ouédraogo, and R. M. Al-Hassan, "The micro determinants of financial inclusion and financial resilience in africa," *African Development Review*, vol. 34, no. 2, pp. 293–306, 2022, doi: https://doi.org/10.1111/1467-8268.12636.
6. A. Das and S. Ghosh, "Financial deregulation and profit efficiency: A nonparametric analysis of indian banks," *Journal of Economics and Business*, vol. 61, no. 6, pp. 509–528, 2009, doi: https://doi.org/10.1016/j.jeconbus.2009.07.003.
7. S. Kumar and R. Gulati, "Did the global financial crisis alter the competitive conditions in the indian banking industry?" *Applied Economics Letters*, vol. 26, no. 10, pp. 857–865, 2019, doi: https://doi.org/10.1080/13504851.2018.1502865.
8. N. Mohan and M. Irfan, "Applying ai & topsis-mcdm tool in evaluating top five private indian bank performances," in *Applications of Block Chain technology and Artificial Intelligence: Lead-ins in Banking, Finance, and Capital Market*. Springer, 2024, pp. 291–303, doi: https://doi.org/10.1007/978-3-031-47324-1_15.
9. G. Caprio, *Banking on crises: expensive lessons from recent financial crises*. World Bank, Development Research Group, Finance, 1998, no. 1979.
10. A. Bhattacharyya, C. K. Lovell, and P. Sahay, "The impact of liberalization on the productive efficiency of indian commercial banks," *European Journal of operational research*, vol. 98, no. 2, pp. 332–345, 1997, doi: https://doi.org/10.1016/S0377-2217(96)00351-7.
11. S. Akhtar, S. N. Azmi, P. A. Khan, A. A. Jan, and Z. Ansari, "Unveiling the financial landscape: analyzing profitability, productivity, and efficiency of banks in an emerging economy using the camels framework and panel analysis," *Cogent Business & Management*, vol. 11, no. 1, p. 2399747, 2024, doi: https://doi.org/10.1080/23311975.2024.2399747.
12. A. Sharma, "Incomplete reform or opportunity: the role of the banking sector in the credit transmission mechanism in india," *Journal of Economic Policy Reform*, vol. 11, no. 4, pp. 273–288, 2008, doi: https://doi.org/10.1080/17487870802567315.
13. B. Eichengreen and P. Gupta, "The financial crisis and indian banks: survival of the fittest?" *Journal of International Money and Finance*, vol. 39, pp. 138–152, 2013, doi: https://doi.org/10.1016/j.jimonfin.2013.06.022.
14. H. Fujii, S. Managi, and R. Matousek, "Indian bank efficiency and productivity changes with undesirable outputs: A disaggregated approach," *Journal of banking & finance*, vol. 38, pp. 41–50, 2014, doi: https://doi.org/10.1016/j.jbankfin.2013.09.022.
15. K. Srinivasan, K. Ramesh, K. Gunasekaran, and K. Sivasubramanian, "Reforms in indian banking sector: a paradigm shift in growth and financial inclusion in india," in *Technology-Driven Business Innovation: Unleashing the Digital Advantage, Volume 1*. Springer, 2024, pp. 433–439, doi: https://doi.org/10.1007/978-3-031-51997-0_36.
16. N. Shikha and I. Kapsis, "Bank crisis management and resolution after svb and credit suisse: Perspectives from india and the european union," *International Insolvency Review*, vol. 33, no. 1, pp. 55–88, 2024, doi: https://doi.org/10.1002/iir.1516.
17. A. Nasim, G. Downing, and M. A. Nasir, "The role of uncertainty, regulatory and economic environment and quantitative tightening in banks' performance," *International Journal of Finance & Economics*, vol. 31, no. 1, pp. 46–69, 2026, doi: https://doi.org/10.1002/ijfe.3128.
18. K. D. Hanumantu, V. Worlikar, and S. Narayanaswami, "The punjab national bank scam: Ethics versus robust processes," *Journal of Public Affairs*, vol. 19, no. 4, p. e1952, 2019, doi: https://doi.org/10.1002/pa.1952.
19. J. Truby, R. Brown, and A. Dahdal, "Banking on ai: mandating a proactive approach to ai regulation in the financial sector," *Law and Financial Markets Review*, vol. 14, no. 2, pp. 110–120, 2020, doi: https://doi.org/10.1080/17521440.2020.1760454.
20. A. Sotoudeh-Anvari, "Root assessment method (ram): A novel multi-criteria decision making method and its applications in sustainability challenges," *Journal of Cleaner Production*, vol. 423, p. 138695, 2023, doi: https://doi.org/10.1016/j.jclepro.2023.138695.
21. ——, "Root assessment method (ram) under neutrosophic environment," in *Neutrosophic Paradigms: Advancements in Decision Making and Statistical Analysis: Neutrosophic Principles for Handling Uncertainty*. Springer, 2025, pp. 123–138, doi: https://doi.org/10.1007/978-3-031-78505-4_6.
22. M. Keshavarz-Ghorabaee, M. Amiri, E. K. Zavadskas, Z. Turskis, and J. Antucheviciene, "Determination of objective weights using a new method based on the removal effects of criteria (merec)," *Symmetry*, vol. 13, no. 4, p. 525, 2021, doi: https://doi.org/10.3390/sym13040525.

23. M. S. Tahir and D. W. Richards, "A systematic literature review of financial resilience: antecedents, consequences and future research agenda," *Journal of Financial Regulation and Compliance*, 2025, doi: https://doi.org/10.1108/JFRC-10-2024-0204.
24. L. Klapper and A. Lusardi, "Financial literacy and financial resilience: Evidence from around the world," *Financial Management*, vol. 49, no. 3, pp. 589–614, 2020, doi: https://doi.org/10.1111/fima.12283.
25. A. Lusardi, A. Hasler, and P. J. Yakoboski, "Building up financial literacy and financial resilience," *Mind & Society*, vol. 20, no. 2, pp. 181–187, 2021, doi: https://doi.org/10.1007/s11299-020-00246-0.
26. C. Jansson, "Financial resilience: the role of financial balance, profitability, and ownership," in *The Resilience Framework: Organizing for Sustained Viability*. Springer, 2017, pp. 111–131, doi: https://doi.org/10.1007/978-981-10-5314-6_7.
27. A. Sreenivasan and M. Suresh, "Readiness of financial resilience in start-ups," *Journal of Safety Science and Resilience*, vol. 4, no. 3, pp. 241–252, 2023, doi: https://doi.org/10.1016/j.jnlssr.2023.02.004.
28. R. L. Clark and O. S. Mitchell, "Americans' financial resilience during the pandemic," *Financial Planning Review*, vol. 5, no. 2-3, p. e1140, 2022, doi: https://doi.org/10.1002/cfp2.1140.
29. G. M. Markman and M. Venzin, "Resilience: Lessons from banks that have braved the economic crisis—and from those that have not," *International Business Review*, vol. 23, no. 6, pp. 1096–1107, 2014, doi: https://doi.org/10.1016/j.ibusrev.2014.06.013.
30. G. O. Danisman, E. Demir, and A. Zaremba, "Financial resilience to the covid-19 pandemic: The role of banking market structure," *Applied Economics*, vol. 53, no. 39, pp. 4481–4504, 2021, doi: https://doi.org/10.1080/00036846.2021.1904118.
31. T. Papadimitriou, P. Gogas, and A. Agrapetidou, "The resilience of the us banking system," *International Journal of Finance & Economics*, vol. 27, no. 3, pp. 2819–2835, 2022, doi: https://doi.org/10.1002/ijfe.2300.
32. C. Cantú, R. Lobato, C. López, and F. López-Gallo, "A loan-level analysis of financial resilience in mexico," *Journal of Banking & Finance*, vol. 135, p. 105951, 2022, doi: https://doi.org/10.1016/j.jbankfin.2020.105951.
33. S. Alam, S. K. Das, U. R. Dipa, and S. Z. Hossain, "Predicting financial distress through ownership pattern: dynamics of financial resilience of bangladesh," *Future Business Journal*, vol. 10, no. 1, p. 91, 2024, doi: https://doi.org/10.1186/s43093-024-00379-5.
34. E. Daadmehr, "Workplace sustainability or financial resilience? composite-financial resilience index," *Risk management*, vol. 26, no. 2, p. 7, 2024, doi: https://doi.org/10.1057/s41283-023-00139-9.
35. Y. Chen and C. Sun, "A new method for measuring financial resilience," *Economics Letters*, vol. 242, p. 111883, 2024, doi: https://doi.org/10.1016/j.econlet.2024.111883.
36. E. K. Zavadskas, Z. Turskis, J. Antucheviciene, and A. Zakarevicius, "Optimization of weighted aggregated sum product assessment," *Elektronika ir elektrotechnika*, vol. 122, no. 6, pp. 3–6, 2012, doi: https://doi.org/10.5755/j01.eee.122.6.1810.
37. S. Gupta, M. Mathew, S. Gupta, and V. Dawar, "Benchmarking the private sector banks in india using mcdm approach," *Journal of Public Affairs*, vol. 21, no. 2, p. e2409, 2021, doi: https://doi.org/10.1002/pa.2409.
38. S. Gupta, M. Mathew, G. Syal, and J. Jain, "A hybrid mcdm approach for evaluating the financial performance of public sector banks in india," *International Journal of Business Excellence*, vol. 24, no. 4, pp. 481–501, 2021, doi: https://doi.org/10.1504/IJBEX.2021.117648.
39. P.-H. Nguyen, J.-F. Tsai, Y.-C. Hu, and G. V. Ajay Kumar, "A hybrid method of mcdm for evaluating financial performance of vietnamese commercial banks under covid-19 impacts," in *Shifting economic, financial and banking paradigm: New systems to encounter COVID-19*. Springer, 2021, pp. 23–45, doi: https://doi.org/10.1007/978-3-030-79610-5_2.
40. A. Karbassi Yazdi, C. Spulbar, T. Hanne, and R. Birau, "Ranking performance indicators related to banking by using hybrid multicriteria methods in an uncertain environment: A case study for iran under covid-19 conditions," *Systems Science & Control Engineering*, vol. 10, no. 1, pp. 166–180, 2022, doi: https://doi.org/10.1080/21642583.2022.2052996.
41. U. Ünlü, N. Yalçın, and N. Avşarlıgil, "Analysis of efficiency and productivity of commercial banks in turkey pre-and during covid-19 with an integrated mcdm approach," *Mathematics*, vol. 10, no. 13, p. 2300, 2022, doi: https://doi.org/10.3390/math10132300.
42. Y. A. Ünvan and C. Ergenç, "Financial performance analysis with the fuzzy copras and entropy-copras approaches," *Computational Economics*, vol. 59, no. 4, pp. 1577–1605, 2022, doi: https://doi.org/10.1007/s10614-021-10143-4.
43. P. Wanke, M. A. K. Azad, A. K. Yazdi, F. R. Birau, and C. M. Spulbar, "Revisiting camels rating system and the performance of asean banks: a comprehensive mcdm/z-numbers approach," *IEEE Access*, vol. 10, pp. 54 098–54 109, 2022, doi: https://doi.org/10.1109/ACCESS.2022.3171339.
44. H. R. Sama, S. V. K. Kosuri, and S. Kalvakolanu, "Evaluating and ranking the indian private sector banks—a multi-criteria decision-making approach," *Journal of Public Affairs*, vol. 22, no. 2, p. e2419, 2022, doi: https://doi.org/10.1002/pa.2419.
45. P. Kumar and D. Sharma, "Prioritising the financial performance of indian private sector banks by a hybrid mcdm approach," *International Journal of Process Management and Benchmarking*, vol. 16, no. 4, pp. 490–511, 2024, doi: https://doi.org/10.1504/IJPMB.2024.137145.
46. Ö. Karadağ Ak, A. Hazar, and Ş. Babuşcu, "Evaluation of the financial performance of development and investment banks with entropy-based aras method," *Macroeconomics and Finance in Emerging Market Economies*, vol. 18, no. 2, pp. 441–461, 2025, doi: https://doi.org/10.1080/17520843.2022.2035523.
47. Ö. Işık, M. Shabir, G. Demir, A. Puska, and D. Pamucar, "A hybrid framework for assessing pakistani commercial bank performance using multi-criteria decision-making," *Financial Innovation*, vol. 11, no. 1, p. 38, 2025, doi: https://doi.org/10.1186/s40854-024-00728-x.
48. J. Zahedi, M. Salehi, and M. Moradi, "Identifying and classifying the contributing factors to financial resilience," *foresight*, vol. 24, no. 2, pp. 177–194, 2022, doi: https://doi.org/10.1108/FS-10-2020-0102.
49. Z. Liu, J.-K. Chen, and J. J. Xiao, "Financial resilience: a scoping review, conceptual synthesis and

theoretical framework," *International Journal of Bank Marketing*, vol. 43, no. 7, pp. 1541–1576, 2025, doi: https://doi.org/10.1108/IJBM-12-2024-0735.
50. T. Jokipii and A. Milne, "Bank capital buffer and risk adjustment decisions," *Journal of Financial Stability*, vol. 7, no. 3, pp. 165–178, 2011, doi: https://doi.org/10.1016/j.jfs.2010.02.002.
51. D. Corbae and P. D'Erasmo, "Capital buffers in a quantitative model of banking industry dynamics," *Econometrica*, vol. 89, no. 6, pp. 2975–3023, 2021, doi: https://doi.org/10.3982/ECTA16930.
52. C. D. Carroll, R. E. Hall, and S. P. Zeldes, "The buffer-stock theory of saving: Some macroeconomic evidence," *Brookings papers on economic activity*, vol. 1992, no. 2, pp. 61–156, 1992, doi: https://doi.org/10.2307/2534582.
53. D. J. Teece, G. Pisano, and A. Shuen, "Dynamic capabilities and strategic management," *Strategic management journal*, vol. 18, no. 7, pp. 509–533, 1997, doi: https://doi.org/10.1002/(SICI)1097-0266(199708).
54. M. M. H. Chowdhury and M. Quaddus, "Supply chain resilience: conceptualization and scale development using dynamic capability theory," *International journal of production economics*, vol. 188, pp. 185–204, 2017, doi: https://doi.org/10.1016/j.ijpe.2017.03.020.
55. M. Doumpos and C. Zopounidis, "A multicriteria decision support system for bank rating," *Decision support systems*, vol. 50, no. 1, pp. 55–63, 2010, doi: https://doi.org/10.1016/j.dss.2010.07.002.
56. R. Ginevičius and A. Podviezko, "The evaluation of financial stability and soundness of lithuanian banks," *Economic research-Ekonomska istraživanja*, vol. 26, no. 2, pp. 191–208, 2013, doi: https://doi.org/10.1080/1331677X.2013.11517616.
57. B. Gavurova, J. Belas, K. Kocisova, and T. Kliestik, "Comparison of selected methods for performance evaluation of czech and slovak commercial banks," *Journal of Business Economics and Management*, vol. 18, no. 5, pp. 852–876, 2017, doi: https://doi.org/10.3846/16111699.2017.1371637.
58. M. Dash, "A model for bank performance measurement integrating multivariate factor structure with multi-criteria promethee methodology," *Asian Journal of Finance & Accounting*, vol. 9, no. 1, pp. 310–332, 2017, doi: https://doi.org/10.5296/ajfa.v9i1.11073.
59. C. Ayadurai and R. Eskandari, "Bank soundness: a pls-sem approach," in *Partial least squares structural equation modeling: Recent advances in banking and finance*. Springer, 2018, pp. 31–52, doi: https://doi.org/10.1007/978-3-319-71691-6_2.
60. I. Marjanović and Ž. Popović, "Mcdm approach for assessment of financial performance of serbian banks," in *Business performance and financial institutions in Europe: Business models and value creation across European industries*. Springer, 2020, pp. 71–90, doi: https://doi.org/10.1007/978-3-030-57517-5_5.
61. J. Reig-Mullor, J. M. Brotons-Martinez, and M. E. Sansalvador-Selles, "A novel approach to improve the bank ranking process: an empirical study in spain," *Journal of Intelligent & Fuzzy Systems*, vol. 38, no. 5, pp. 5323–5331, 2020, doi: https://doi.org/10.3233/JIFS-179626.
62. M. A. Bushashe, "Determinants of private banks performance in ethiopia: A partial least square structural equation model analysis (pls-sem)," *Cogent Business & Management*, vol. 10, no. 1, p. 2174246, 2023, doi: https://doi.org/10.1080/23311975.2023.2174246.
63. V. Sharma, M. Gupta, and K. Jangir, "Exploring the impact of risk factors on profitability in commercial banking in india: a pls-sem analysis approach," 2024, doi: https://doi.org/10.1108/978-1-83753-734-120241007.
64. S. S. Goswami, S. K. Mohanty, and D. K. Behera, "Selection of a green renewable energy source in india with the help of merec integrated piv mcdm tool," *Materials today: proceedings*, vol. 52, pp. 1153–1160, 2022, doi: https://doi.org/10.1016/j.matpr.2021.11.019.
65. S. Kumar, P. Ahijith Kumar, K. Bharati, L. Patnaik, S. Ranjan Maity, and M. Łepicka, "Coating material selection for bulk metal forming dies: A merec-integrated approach with multiple mcdm methods," *International Journal on Interactive Design and Manufacturing (IJIDeM)*, vol. 19, no. 6, pp. 4055–4070, 2025, doi: https://doi.org/10.1007/s12008-024-01983-z.
66. I. M. Hezam, A. M. Ali, K. Sallam, I. A. Hameed, A. Foul, and M. Abdel-Basset, "An extension of root assessment method (ram) under spherical fuzzy framework for optimal selection of electricity production technologies toward sustainability: a case study," *International Journal of Energy Research*, vol. 2024, no. 1, p. 7985867, 2024, doi: https://doi.org/10.1155/2024/7985867.
67. T. Van Dua, D. D. Trung, N. T. P. Giang, and D. Van Duc, "Comparison of two methods: Ram and aroman," in *International Conference on Sustainability and Emerging Technologies for Smart Manufacturing*. Springer, 2024, pp. 727–735, doi: https://doi.org/10.1007/978-981-97-7083-0_73.
68. A. Shekhovtsov, M. Gandorc, R. Pawlakb, and W. Sałabuna, "A novel rancom-ram-based framework for city assessment based on cost of living," *Procedia Computer Science*, vol. 270, pp. 5776–5786, 2025, doi: https://doi.org/10.1016/j.procs.2025.10.046.



BenchCouncil Transactions on Benchmarks, Standards and Evaluations, 2026

DOI: https://doi.org/10.66834/byjwsd23
Research Article

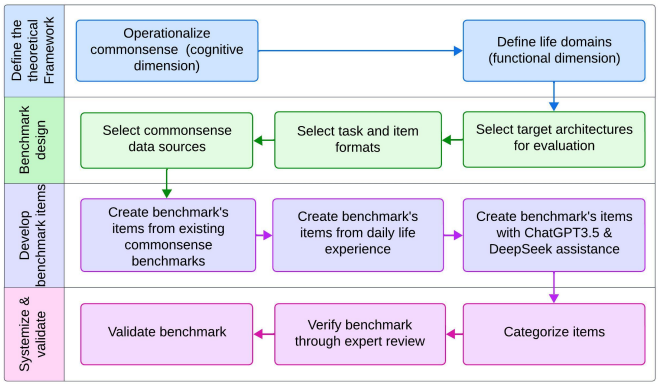
RESEARCH ARTICLE

JAMAL: A Multidimensional Benchmark for Arabic Commonsense Reasoning Across Life-Domains and Cognitive Axes

Basma Sayah,[1,*] Attia Nehar,[1,2] Hadda Cherroun,[1] Slimane Bellaouar[3] and Firoj Alam[4]

[1]Laboratoire d'Informatique et de Mathématiques (LIM), Amar Telidji University, Laghouat, Algeria , [2]Computer Science Department, Ziane Achour University of Djelfa, Djelfa, Algeria , [3]Laboratoire de Mathématiques et Sciences Appliquées (LMSA), Université de Ghardaia, Ghardaia, Algeria and [4]Qatar Computing Research Institute (QCRI), Hamad Bin Khalifa University, Doha, Qatar
*Corresponding author: b.sayah@lagh-univ.dz
Received on 10 January 2026; Accepted on 21 May 2026

Graphical Abstract



Highlights

- JAMAL, a novel language-agnostic commonsense framework instantiated as a concrete Arabic benchmark.
- Establishes a functional dimension covering 56 life-domain categories.
- Defines a cognitive dimension encompassing everyday situations, general knowledge, and problem-solving.
- Introduces a cultural grounding categorization of items into universal, western/global, and Arabic-specific knowledge for controlled analysis.
- Conducts a fine-grained evaluation of five language models across the three axes, revealing behavioral error patterns.
- Publicly releases the dataset to foster Arabic NLP development and support future multilingual extensions.

Abstract

Commonsense is a broad and multifaceted concept, making its evaluation a persistent challenge in natural language processing (NLP). This paper introduces **JAMAL** (Arabic for "Camel"), a multidimensional framework and benchmark for Arabic commonsense reasoning. JAMAL is structured along three complementary axes: (i) a life-domain axis comprising a taxonomy of 56 functional categories informed by the World Health Organization's International Classification of Functioning, Disability, and Health (ICF), capturing diverse aspects of daily human experience; (ii) a cognitive axis organizing commonsense into three reasoning types: everyday situations, general knowledge, and problem-solving; and (iii) a cultural grounding axis distinguishing between universal, western/global, and Arabic-specific knowledge. To operationalize this framework, benchmark items are constructed using psycholinguistically inspired principles of constrained contextual prediction. We evaluate five Arabic language models using JAMAL and observe consistent differences in their performance across all axes. Notably, FANAR-27B achieves the strongest overall results among all evaluated models, outperforming FANAR-9B and smaller baselines. Overall, **JAMAL** provides a structured and interpretable benchmark for evaluating commonsense reasoning in Arabic, supporting the development of more robust language models through systematic analysis of their behavioral limitations.

Key words: Commonsense Reasoning, Arabic NLP, Language Model Evaluation, Psycholinguistically Grounded Benchmarking, WHO-ICF Framework

1. Introduction

Neural language models have made remarkable strides in recent years, achieving strong performance across a wide range of NLP tasks, including text generation, sentiment analysis, and machine translation [1]. In particular, large-scale transformer-based models such as GPT, Llama, Qwen, and Gemini can generate fluent and coherent text [2–5].

As these models grow in capability and complexity, rigorous evaluation becomes increasingly important. Standard benchmark suites probe abilities such as commonsense reasoning, general knowledge, and reading comprehension, including HellaSWAG [6], MMLU [7], and RACE [8]. However, most of these benchmarks are designed primarily for English, limiting their direct applicability to other languages without adaptation.

In Arabic NLP, evaluation has advanced in recent years through general benchmarking efforts [9, 10], dedicated resources such as ArabicMMLU [11] and AraDiCE [12], and leaderboards such as BALSAM [13]. However, benchmarks targeting other specific capabilities remain limited [14].

Among these capabilities, commonsense reasoning is particularly important, as it underpins effective understanding and interaction with the world [15]. Despite its importance, existing Arabic commonsense evaluation resources remain limited in both scope and granularity. Early efforts, such as the translated *Is This Sentence Valid?* dataset [16], provide only coarse, task-level evaluation and do not capture the multifaceted nature of commonsense reasoning.

At the same time, the emergence of Arabic large language models—including Jais [17], ALLaM [18], and Fanar [19] has increased the need for more structured and diagnostic evaluation frameworks. These models differ in training data composition, dialectal coverage, and intended use cases, and their capabilities are not fully reflected by existing general-purpose benchmarks. This motivates the need for structured evaluation frameworks that provide deeper insight into model behavior.

To address these challenges, we introduce **JAMAL**[1], a multidimensional framework and Arabic benchmark for evaluating commonsense reasoning in language models.

JAMAL is structured along three complementary axes. The first is a life-domain axis comprising 56 functional categories informed by the World Health Organization's International Classification of Functioning, Disability, and Health (ICF) [20], enabling systematic coverage of everyday human experiences. The second is a cognitive axis that organizes commonsense into three reasoning types: everyday situations, general knowledge, and problem-solving. The third is a cultural grounding axis distinguishing between universal, western/global, and Arabic-specific knowledge.

Together, these axes enable fine-grained and interpretable evaluation of model performance across functional domains, reasoning types, and cultural contexts.

This work makes four contributions: (1) a language-agnostic multidimensional framework for commonsense evaluation; (2) an Arabic benchmark grounded in functional, cognitive, and cultural axes; (3) a comprehensive empirical analysis of model behavior across these dimensions; and (4) diagnostic insights revealing systematic strengths and weaknesses across five Arabic language models.

[1] The name *JAMAL* is a wordplay: in Arabic, *jamal* means "Camel," a culturally salient symbol, and it is also phonetically related to *jumal* ("sentences"), reflecting the benchmark's sentence-based design.

The remainder of this paper is organized as follows. Section 2 reviews related work on commonsense reasoning evaluation, covering global benchmarks, multilingual efforts, and Arabic-specific resources. Section 3 describes the design and construction of JAMAL. Section 4 presents the evaluation of five Arabic language models using JAMAL. Finally, Section 5 discusses conclusions, limitations, and future directions.

2. Related Work

This section reviews the evolution of global and Arabic commonsense evaluation benchmarks to situate the proposed JAMAL within the existing literature.

2.1. Global trends in commonsense evaluation

At the global level, commonsense reasoning (CSR) evaluation has evolved through a series of influential benchmarks, reflecting a shift from performance-oriented assessment toward more structured and diverse methodologies. Early large-scale datasets, such as SWAG (2018) [21], CommonsenseQA (2019) [22], and HellaSwag (2019) [6], predominantly adopt multiple-choice question answering formats, where commonsense is treated as a single unified capability and evaluated using accuracy-based metrics.

In the following years, evaluation was extended to more specialized reasoning domains. Social IQA (2019) [23] focuses on social and emotional reasoning, while PIQA (2020) [24] targets physical reasoning about everyday object use, material properties, and affordances. This shift was motivated by the need to better capture different facets of commonsense beyond general-purpose reasoning. However, most benchmarks still rely on answer selection as the primary evaluation paradigm.

In parallel, multilingual benchmarks were introduced to study cross-lingual transfer of commonsense knowledge. X-CODAH (2019) [25] and X-CSQA (2021) [26] extend existing datasets to multiple languages, typically via translation-based approaches, enabling evaluation of multilingual generalization.

More recently, research has moved toward more structured and interpretable evaluation frameworks. TG-CSR (2023) [27] introduces a theory-driven approach that decomposes commonsense reasoning into nine interpretable dimensions, such as temporal, spatial, and causal reasoning, enabling more fine-grained analysis of model behavior.

A further trend is the shift toward generative and reasoning-centric evaluation. For example, ExplaGraphs (2021) [28] requires models to generate structured explanations to justify predictions, while emerging frameworks such as SCoRE (2025) [29] emphasize multi-hop reasoning and reasoning-chain evaluation. Overall, these developments reflect a movement from static accuracy-based benchmarks toward more interpretable and reasoning-aware evaluation paradigms.

Despite these advances, a tension remains between scalability and interpretability: large-scale benchmarks offer broad coverage and simple evaluation, while structured approaches provide deeper insights but are less scalable.

2.2. Commonsense evaluation for Arabic language

Research on Arabic commonsense reasoning has produced a growing number of benchmarks, establishing important foundations for Arabic-centric evaluation while also revealing challenges in achieving structured and fine-grained assessment.

Early benchmarks, such as the Arabic Commonsense Dataset (ArCD) (2019) [30], Is This Sentence Valid?

Benchmark	Year	Lang.	Methodology	Taxonomy / Structure	Evaluation
SWAG [21]	2018	EN	MCQ (sentence completion)	None	Accuracy
CSQA [22]	2019	EN	MCQ (knowledge-based)	ConceptNet relations	Accuracy
HellaSwag [6]	2019	EN	Adversarial MCQ	None	Accuracy
Social IQA [23]	2019	EN	MCQ (social reasoning)	Social scenarios	Accuracy
PIQA [24]	2020	EN	MCQ (physical reasoning)	Physical interactions	Accuracy
X-CODAH [25]	2019	Multi	Multilingual MCQ	Translated dataset	Accuracy
X-CSQA [26]	2021	Multi	Multilingual MCQ	Translated taxonomy	Accuracy
ExplaGraphs [28]	2021	EN	Generative explanation	Argument graphs	NLI + explanation quality + human eval
TG-CSR [27]	2023	EN	Theory-driven MCQ	9 reasoning dimensions	Accuracy per dimension
SCoRE [29]	2025	EN	Multi-hop reasoning	Scenario-based logic	CoT audit + human eval
ArCD [30]	2019	AR	MCQ (Wikipedia-based)	None	Accuracy
Is This Sentence Valid? [16]	2020	AR	Sentence classification	Binary validity	Accuracy
Arabic Winograd [31]	2020	AR	Coreference resolution	Pronoun resolution	Accuracy
ArabicSense [32]	2025	AR	MCQ + generation	Implicit / synthetic	Accuracy, F1, BERTScore
Commonsense in Arab Culture [33]	2025	AR	MCQ + completion	12-domain taxonomy	Accuracy
JAMAL	**2026**	**AR**	**CPARG-based cloze-style text completion (psycholinguistically motivated); hybrid construction (manual curation + semi-automatic generation)**	**Three-axis structure: cognitive (3 reasoning types), life-domain (56 categories), and cultural grounding (3 categories)**	**Accuracy across axes**

Table 1. Comparison of existing global and Arabic commonsense reasoning benchmarks across language, methodology, taxonomy, and evaluation metrics, including the proposed JAMAL benchmark.

(2020) [16], and the Arabic Winograd Dataset (2020) [31], introduced initial testbeds for Arabic commonsense evaluation. These datasets mainly rely on multiple-choice or classification formats and treat commonsense as a general and undifferentiated capability.

More recent work has attempted to enrich both task design and evaluation objectives. ArabicSense (2025) [32] incorporates both classification and explanation generation, using metrics such as accuracy, F1, and BERTScore, although it relies heavily on synthetic data. Commonsense Reasoning in Arab Culture (2025) [33] introduces a culturally grounded taxonomy covering 12 daily life domains and 54 subtopics, combining multiple-choice and sentence completion tasks to evaluate reasoning across Arab cultural contexts. While this improves coverage, the taxonomy is largely derived from corpus-driven topic modeling with manual refinement, rather than an explicit cognitive or functional framework.

To address these limitations, there is a growing need for theory-driven benchmarks that implement structured taxonomies grounded in established cognitive, psychological, or behavioural models. Such benchmarks should also ensure careful item construction to avoid synthetic artifacts or superficial lexical cues, instead capturing meaningful reasoning processes.

Motivated by these limitations, we introduce JAMAL, a language-agnostic framework designed around a taxonomy spanning functional, cognitive, and cultural dimensions. Built through a controlled human-in-the-loop construction process, JAMAL enables fine-grained evaluation of commonsense reasoning across multiple complementary axes, supporting more systematic analysis of model behavior than existing Arabic benchmarks. Table 1 summarizes the key characteristics of global and Arabic benchmarks and situates JAMAL within the broader evaluation landscape.

3. The JAMAL Benchmark: Design, Development and Validation

This section describes the design and construction of JAMAL, a language-agnostic framework for fine-grained, multidimensional commonsense evaluation, and its Arabic instantiation as a concrete benchmark dataset. The overall process is illustrated in Figure 1.

3.1. Define the theoretical framework

In this first step, we aim to establish a theoretical foundation for understanding commonsense, in order to evaluate it effectively.

3.1.1. Operationalize commonsense (Cognitive dimension)

Since the goal was to build an effective benchmark for evaluating the commonsense knowledge of language models and to assess the extent to which this knowledge is acquired during training, it was essential to establish a clear and precise definition of commonsense to delineate what constitutes commonsense and what does not. After reviewing several definitions, we adopted one that explicitly distinguishes commonsense from domain-specific expertise: *"Commonsense is practical good sense gained through life experience, not through specialized study"* [34].

To operationalize this concept, we identified key themes from the literature. Ilievski et al. [35] emphasize the dimension of everyday situational knowledge, which enables navigation of routine scenarios. In contrast, Lenat [36] and Whiting et al. [37] treat commonsense as a body of general factual knowledge about the world. Complementing these, Newell and Simon [38] link it to the problem-solving processes used to address everyday challenges.

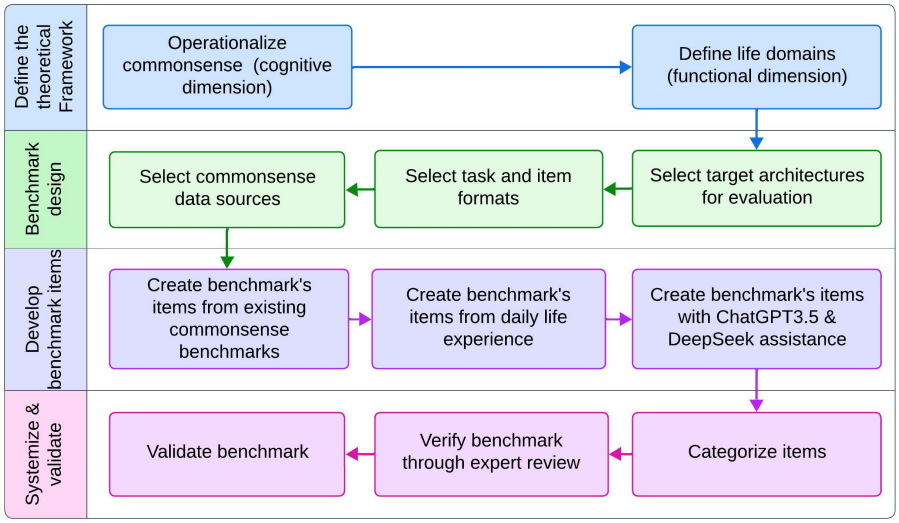


Figure 1. JAMAL Construction Process: From defining commonsense and its dimensions to benchmark validation.

Synthesizing these perspectives, we derived three core categories of commonsense: *(i)* everyday situations, *(ii)* general knowledge, and *(iii)* problem-solving.

The dimension of **everyday situations** encompasses the ability to understand typical scenarios, predict events based on prior experience [39], make sound judgments, and respond appropriately to common occurrences [40]. For example, understanding that people typically stand in a queue to wait for their turn or that one usually checks if an appliance is plugged in when it does not turn on. The **general knowledge** dimension refers to the factual and conceptual understanding that individuals typically possess, including basic facts, causal relationships, and general principles (e.g., "the sky is blue" [36]). This category also includes knowledge about physical objects and their properties, such as understanding that *an egg consists of a yolk, egg white, and shell* [37]. The **problem-solving** dimension involves applying logical reasoning and cognitive skills to solve structured tasks and address real-world challenges that require decision-making and pattern recognition [38]. This includes scenarios such as prioritizing which tasks to complete first when facing multiple deadlines. Furthermore, this dimension is shaped by interactions with both the environment and social contexts [41]. While these three categories capture the cognitive aspects of commonsense reasoning, its evaluation also requires grounding it in diverse areas of everyday life. To this end, we introduce a complementary life-domain dimension, described in the following subsection.

3.1.2. Define Life Domains (Functional Dimension)

We adopt the *International Classification of Functioning, Disability and Health (ICF)* as the conceptual foundation for defining the functional dimension of JAMAL, leveraging its comprehensive coverage of human activities, participation, and environmental context. However, our goal is not to reproduce the clinical taxonomy, but to derive a simplified and functionally meaningful set of life domains suitable for commonsense reasoning evaluation.

The taxonomy is derived by clustering related ICF components (primarily d- and e-codes) into higher-level functional domains that reflect how humans organize everyday experience. Rather than enforcing a one-to-one mapping between ICF codes and items, we aggregate multiple codes when they correspond to the same commonsense reasoning context.

For example, mobility-related activities (d4) inform *Transportation and Movement*, interpersonal interactions (d7) inform *Personal and Social Life*, and leisure-related activities (d920) are distributed across *Games, Sports, and Recreation*, *Arts, Music, and Creativity*, and *Fiction and Entertainment*. Environmental factors such as products, nature, and animals (e1, e2, e245) are integrated into domains describing physical and ecological context.

This procedure yields 15 top-level life domains covering social, physical, environmental, cognitive, occupational, and recreational aspects of human life, further decomposed into 56 subdomains (see Table A).

We introduce three deliberate simplifications relative to the original ICF: (i) aggregating multiple activity codes that correspond to the same everyday commonsense reasoning context into unified functional domains, (ii) elevating perceptual attributes such as *Colors, Shapes, and Forms* from sensory functions to a standalone commonsense domain, and (iii) separating goal-directed activities (*Human Actions and behaviours*) from institutional contexts (*Work and Productivity*) to better reflect differences in reasoning demands.

Overall, the resulting taxonomy preserves the breadth of the ICF while providing an evaluation-oriented organization of everyday human experiences for commonsense reasoning tasks.

3.2. JAMAL Design

This design phase translates the theoretical framework into practical specifications. We first identify the target language models for evaluation, which informs the selection of suitable benchmark item formats. We then select foundational knowledge sources aligned with the functional and cognitive dimensions of the framework.

3.2.1. Select target architectures for evaluation

Given the diversity of available architectures, we selected two main categories of language models for evaluation: **base language models** and **instruction-tuned models**.

Base language models are pre-trained on large, general-purpose corpora using objectives such as masked or next-token prediction. They typically require task-specific fine-tuning to achieve strong performance, as exemplified by models like BERT [42] and GPT [43].

instruction-tuned models are large language models that undergo additional fine-tuning to follow natural language instructions. This paradigm, popularized by models such as ChatGPT [2], enables zero-shot and few-shot task execution without the need for explicit task-specific fine-tuning.

3.2.2. Select task and item formats

We adopted the CPARG assessment format, a psycholinguistically grounded approach employed in *"What BERT Is Not?"* [44], which was originally introduced in human studies by Federmeier and Kutas [45]. In this format, each item presents a two-sentence context in which the task is to predict the final word of the second sentence. The task requires using implicit cues from the first sentence to infer this missing word through commonsense reasoning.

An example of a context in the CPARG format is shown below:

- The child blew out the candles. Everyone shouted happy ____.

In this example, the target word to be predicted is *birthday*. Solving this item relies on commonsense knowledge typically shared through a familiar social script of blowing out candles at a birthday celebration.

The CPARG format allows for the evaluation of the internal commonsense knowledge that language models acquire during training. Since the target word in each item is not explicitly stated, models cannot rely on simple text matching or selecting from predefined options. Instead, they must draw on internal reasoning and inference to predict the correct word.

To ensure proper alignment with the CPARG format, we established the following criteria for creating the benchmark items. In line with CPARG terminology, we refer to these items as *contexts* :

- Contexts must reflect commonsense rather than specialized knowledge, following the established definition 3.1.1.

  Example of specialized knowledge (invalid):

  Sentence 1: The patient's echocardiogram revealed severe mitral valve regurgitation.

  Sentence 2: The cardiologist decided the best course of action was to perform a [blank].

  Target word: *annuloplasty*
- Each context must consist of exactly two sentences.
- The target word to be predicted must not be explicitly mentioned in the context.

  Example of an explicitly mentioned target word in the first sentence (invalid):

  Sentence 1: He couldn't find his keys anywhere.

  Sentence 2: After searching for an hour, he finally found the missing [blank].

  Target word: *keys*
- The dataset must cover a diverse range of commonsense scenarios representing different aspects of daily life.

3.2.3. Select commonsense data sources

To construct commonsense items in the CPARG format, we required source material that could be reformulated into two-sentence contexts. We drew upon three main sources: existing commonsense benchmarks, scenarios inspired by our everyday experiences, and text generated by ChatGPT-3.5 and DeepSeek.

This multi-source approach was strategically chosen to leverage the unique advantages of each while mitigating their inherent biases. Existing benchmarks provide a validated foundation but risk inheriting cultural biases and data contamination from their original creation. Scenarios from daily life introduce crucial realism and a human perspective, yet they are limited in scalability and can reflect the subjectivity of the researchers' own experiences. Finally, LLM generated text ensures scalability and diversity but may amplify social biases present in the models' training data or produce superficially fluent but conceptually shallow items. By combining these sources and refining them manually, JAMAL aims to encompass a broad spectrum of commonsense knowledge, counterbalancing the weaknesses of any single source to create a more robust and nuanced evaluation set.

3.3. Develop benchmark items

In this section, we describe the process of creating and curating items from diverse sources in accordance with the prescribed CPARG format.

3.3.1. Create Benchmark's items from existing commonsense benchmarks

We searched for existing benchmarks using the three categories of the cognitive dimension of commonsense as keywords, focusing on assessments from both human and machine studies that incorporated these categories. For everyday situations, we identified the HellaSWAG [6] and Winogrande [46] benchmarks. HellaSWAG contains short narrative scripts drawn from video captions. Winogrande focuses on pronoun resolution in multiple-choice questions, testing comprehension of context and references. For general knowledge, we selected the "Is This Sentence Valid?" benchmark [16], which evaluates the ability to determine the factual and logical validity of statements. For problem-solving, we selected the Cornell Conditional Reasoning Test [47], a standardized assessment from cognitive psychology designed to evaluate logical and conditional reasoning in humans.Figure 2 illustrates the selected benchmarks.

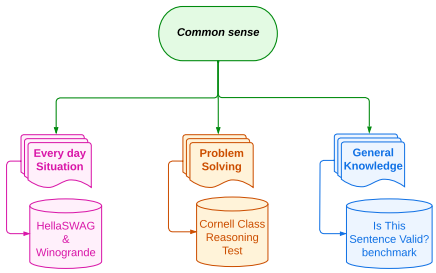


Figure 2. Overview of selected commonsense benchmarks categorized into three core cognitive dimensions: everyday Situations, general Knowledge, and problem Solving.

No.	Processing Stage	Context	السياق	Issue
1	Original	He crashed into a brick wall and fell to the ground. As he lifted his wrist, someone's shoe came down on it.	اصطدم بجدار من الطوب وانهار على الأرض. يرفع معصمه، ولكن حذاء شخص ما ينزل عليه.	Unusable context; no clear commonsense inference can be derived.
1	**Discarded**	**Discarded sentence**	**تم الاستبعاد**	**Rejection due to irrelevance.**
2	Original	The person lifts the violin to their chin and prepares. They play a song on the violin.	يرفع الشخص الكمان إلى ذقنه ويستعد. يعزف الشخص أغنية على الكمان.	The word "violin" can serve as the inference target.
2	**Refined context**	**Ahmad lifts his instrument to his chin and gets ready. He plays a sad song on the violin.**	يرفع أحمد آلته إلى ذقنه ويستعد. يقوم بعزف أغنية حزينة على الكمان.	**Removed the target word from the first sentence; added the emotional cue "sad" and the contextual cue "instrument".**
3	Original	A fiery ball throws a person backward. He crashes through a brick wall and lands on the ground.	كرة نارية ترمي شخصًا إلى الوراء. يتم دفعه عبر جدار من الطوب ويهبط على الأرض.	Unrealistic scenario; metaphorical description lacking grounded commonsense interpretation.
3	**Newly crafted**	**His eyes were swollen and bruised. He got into a fight yesterday and received a punch.**	كانت عيناه متورمتين وزرقاوين. دخل في شجار البارحة وتلقى لكمة.	**Inspired by physical harm, we replaced the unrealistic scenario with a grounded everyday situation involving a fight and injury.**

Table 2. Examples of items manually created or refined from existing commonsense benchmarks.

We translated the selected benchmark sentences into Arabic using a Python script and the Google Statistical Machine Translation (SMT) API[2]. Sentences were manually adapted to the CPARG format, discarding many captions, questions, and dialogues that did not reference a specific word or concept. Three main adaptation cases were observed:

- **Unusable sentences:** Could not be modified to CPARG format as no target word could be inferred from context (see sentence (1), Table 2).
- **Easily adaptable sentences:** Required minimal edits, such as removing the final word to designate it as the target (see sentence (2), Table 2).
- **Sentences requiring full rewriting:** Retained only the core idea; new sentences were crafted to fit the CPARG format (see sentence (3), Table 2).

A total of 600 items were adapted. The process was highly selective, except for the Cornell Critical Thinking Test, Level X, whose 65 items were included in full. These items assess reasoning and judgment through everyday scenarios and explicitly aim to evaluate "the ability to apply logical principles to everyday problems" [47], making them well-suited for commonsense evaluation.

3.3.2. Create benchmark's items from daily life experience

To expand the benchmark's scope and realism, we supplemented it with 400 new items inspired by our own everyday experiences. Many of these ideas emerged from reviewing existing benchmarks or from personal observations. Each item was carefully constructed to conform to the CPARG format, resulting in a benchmark of 1,000 contexts. Examples of contexts from our daily life experience are shown in table 3:

3.3.3. Benchmark Item Construction using ChatGPT-3.5 and DeepSeek

We used two LLM-based generation strategies to construct CPARG-format contexts: (i) interactive prompting with ChatGPT-3.5, and (ii) an automated multi-step pipeline using the DeepSeek API.

ChatGPT-3.5 interactive generation:

We first generated CPARG-format contexts using ChatGPT-3.5 through iterative prompt engineering. We experimented with different prompting strategies, ranging from minimal prompts (providing only CPARG examples) to more explicit instructions describing the task, and finally to detailed step-by-step prompts encouraging constraint checking before generation. A representative prompt is:

> "Generate 10 Arabic two-sentence contexts: the first sentence provides a hint about a food item, and the second sentence ends with the target item as the final word. Ensure that the first sentence uniquely identifies the item and that the second sentence reveals the answer only in the final word."

This strategy was not fully error-proof, given the inherent difficulty of implicitly eliciting target words without explicit mention. Common issues included the target word appearing

[2] `https://cloud.google.com/translate`

No.	Example (Source)	Translation
1	كانت الطاولة مغطاة بالغبار. أحضرت ليلى قطعة قماش وبدأت بمسحها.	The table was covered in dust. Leila brought a cloth and started wiping.
2	كان عبد القادر وهشام يحبان التخييم. فور وصولهما إلى وجهتهما بدآ بجمع الحطب.	Abdelkader and Hisham loved camping. As soon as they arrived, they started collecting firewood.
3	كانت بسمة تعاني من ضعف في النظر. نصحتها أمها بأكل الجزر.	Basma suffered from poor eyesight. Her mother advised her to eat carrots.

Table 3. Examples of CPRAG-style contexts derived from daily-life experiences.

No.	Proceesing Stage	Context	السياق	Issue
1	Generated	I drew a new card and remembered the rules. There was no room for mistakes in UNO.	أخذت بطاقة جديدة وتذكرت القواعد. لم يكن هناك مجال للخطأ في اونو.	Too vague, applies to multiple card games.
1	**Refined**	**She looked at the colored cards and had to find a way to get rid of them quickly. There was no room for error in playing UNO.**	نظرت إلى البطاقات الملونة، كان عليها أن تجد طريقة للتخلص منها بسرعة. لم يكن هناك مجال للخطأ في لعب اونو.	**Added gameplay-specific cues (colored cards and card-elimination strategy) to better disambiguate the UNO context.**
2	Generated	He took his racket in hand, realizing that hitting with power and precision was the key to victory. The match was all about exchanging hits on the tennis court.	أخذ مضربه بيده، مدركاً أن الضرب بقوة ودقة هو مفتاح الفوز. كانت المباراة تدور حول تبادل الضربات على ملعب التنس.	Ambiguous terms: racket and court; could refer to other racket sports.
2	**Refined**	**The match was about exchanging the bright green ball with the racket. Every day, he would go to practice on the tennis court.**	كانت المباراة تدور حول تبادل الكرة ذات اللون الأخضر الفاقع بالمضرب. كان كل يوم يذهب للتدرب في ملعب التنس.	**Added a concrete visual cue (green tennis ball) and habitual practice context to better disambiguate the tennis scenario.**
3	Generated	When Ahmed's family prepares a healthy dinner, they like to add dark green leafy vegetables rich in minerals, like spinach.	عندما تعد عائلة أحمد وجبة عشاء صحية، فإنهم يحبون إضافة خضروات ذات أوراق خضراء داكنة وغنية بالمعادن، مثل السبانخ.	Fits multiple vegetables.
3	**Refined**	**Ahmed's family prepared a meal that Popeye always eats to become strong. But Ahmed did not like eating spinach.**	عندما أعدت عائلة أحمد وجبة يأكلها باباي دائماً ليصبح قوياً. لكن أحمد كان لا يحب تناول السبانخ.	**Added strong cultural cue (Popeye reference) to uniquely identify spinach as the intended vegetable.**

Table 4. Examples of CPARG-format contexts generated by ChatGPT-3.5 and their subsequent manual refinement.

before the final position in the sentence, ambiguous or underspecified hints (e.g., "she steamed this green vegetable. . . "), and reduced quality when generating multiple examples simultaneously. These limitations are consistent with known challenges in constraint adherence and hallucination in large language models [48].

All outputs were therefore manually reviewed, and noncompliant instances were corrected to ensure strict adherence to the CPARG format. Examples of this refinement process are shown in Table 4.

DeepSeek API pipeline generation:
To improve scalability and reduce manual interaction, we additionally used the DeepSeek API with an automated multi-step generation pipeline. This pipeline decomposed the task into three stages:

1. Generate target words belonging to functional categories (see Appendix A).
2. Generate short descriptive hint sentences for each target word.

No.	Processing Stage	Context	السياق	Issue
1	Generated	On the slopes of the mountains, tall green plants rose, famous for their sturdy wood. We took souvenir pictures next to the cedar tree.	على سفوح الجبال ارتفعت نباتات شاهقة خضراء طيلة العام، تشتهر بخشبها المتين. أخذنا صورًا تذكارية بجانب شجرة الأرز.	Could refer to other types of trees.
1	**Refined**	**On the slopes of the Lebanese mountains, tall green plants rose, famous for their sturdy wood. We took souvenir pictures next to the cedar tree.**	على سفوح جبال لبنان ارتفعت نباتات شاهقة خضراء طيلة العام، تشتهر بخشبها المتين. أخذنا صورًا تذكارية بجانب شجرة الأرز.	**Added a strong geographic cue (Lebanon), since the cedar tree is a national symbol of the country.**
2	Generated	I wanted to grow plants that tolerate salinity and give sweet fruits near the sea. The expert advised me to plant a palm tree.	أردت زراعة نباتات تتحمل الملوحة وتعطي ثمارًا حلوة بالقرب من البحر. نصحني الخبير بزراعة النخيل.	Ambiguous plant context, as it allows multiple interpretations of salt-tolerant plants.
2	**Refined**	**I wanted to produce dates rich in benefits. I have decided to plant a palm tree.**	أردت إنتاج التمر الغني بالفوائد. قررت زراعة النخيل.	**Added "dates" cue to disambiguate the target.**
3	Generated	In the fertile fields, tall plants were grown with stems full of a sweet-tasting liquid from which a natural sweetener is extracted. To produce natural sugar, The farmers planted a lot of sugarcane.	في الحقول الخصبة، كانت تزرع نباتات طويلة ذات سيقان مليئة بسائل حلو المذاق يستخرج منه المحلي الطبيعي. لانتاج السكر الطبيعي، زرع الفلاحون الكثير من قصب	CPRAG should rely on the full context rather than the second sentence alone.
3	**Refined**	**In the fertile fields, tall plants were grown with stems full of a sweet-tasting liquid from which a natural sweetener is extracted. The farmers planted a lot of sugarcane.**	في الحقول الخصبة، كانت تزرع نباتات طويلة ذات سيقان مليئة بسائل حلو المذاق يستخرج منه المحلي الطبيعي. زرع الفلاحون الكثير من قصب	**Removed the additional cue ("to produce natural sugar") to ensure that inference is based on the full context rather than the second sentence alone.**

Table 5. Examples of CPARG-style contexts generated by the DeepSeek pipeline, together with their subsequent manual refinement.

3. Generate an intermediate sentence that semantically links the hint sentence (first sentence) to the target word, which appears as the final word of the second sentence.

While this structured pipeline improved consistency, manual validation and correction were still required to ensure full compliance with CPARG constraints. Representative examples of refined outputs are shown in Table 5.

Overall, combining ChatGPT-3.5 and DeepSeek-based generation, we constructed a total of 1823 contexts.

It is worth noting that in English translations of the generated examples, the target word may not always appear in final position due to differences in syntactic structure. However, this does not affect the original Arabic instances used in the benchmark, where the target word is strictly constrained to appear as the final token in accordance with the CPARG format.

3.4. Systemize and validate

In this final phase, we systematize the benchmark by labeling each item according to its functional and cognitive category assignments. We then conduct human verification and validation of both the labels and the benchmark items to ensure the reliability and internal consistency of JAMAL.

3.4.1. Categorize items

To enable fine-grained evaluation, each item in JAMAL was labeled with one of the 56 life-domain categories from the functional dimension (Appendix A) and with one or more of the three cognitive branches (everyday situations, general knowledge, and problem-solving). Because the cognitive branches can overlap, items could receive multiple cognitive labels when appropriate. All automatically assigned labels were subsequently verified and corrected by human annotators.

Appendix B details the final distribution of items across categories. Across its 56 subdomains, JAMAL contains between 17 and 54 items per subcategory. This density aligns with established benchmark practices: BIG-bench tasks often contain 10–100 examples per task, CommonsenseQA and HellaSwag include roughly 15–50 examples per category, Social IQA contains approximately 15–50 questions per subcategory, and psycholinguistic diagnostic suites such as CPARG consist of 102 items distributed across all categories.

Each item was also annotated for cultural grounding to enable post-hoc analysis of model performance across different types of cultural knowledge. Items were labeled into three categories: *universal* (knowledge shared across cultures), *western/global* (concepts commonly encountered in mainstream media), and *Arabic-specific* (knowledge grounded in Arabic cultural contexts, including traditions and local practices), as detailed in Section 4.3.2.

3.4.2. Verify benchmark through expert review

The verification process consisted of two stages. First, one of the authors reviewed the items to ensure compliance with the benchmark design and corrected any inconsistencies. Second, an external native Arabic speaker independently evaluated the items and flagged contexts that did not meet the CPARG requirements.

The reviewer received detailed written instructions (see Appendix C), which included:

- Carefully examining each context in the benchmark.
- Identifying contexts that violated the established requirements.
- Providing justification for each flagged context.
- Noting ambiguous or unclear formulations.

During the verification process, 83 contexts (4.6%) were flagged as incorrect and categorized into four error types:

- **Target word already mentioned in the context (37 out of 1823):** This was the largest error category, comprising 44.6%of all errors.
- **Multiple possible target words (16 out of 1823):** This category includes cases where more than one target word could be inferred, comprising 19.3%of all errors.
- **Incorrect target word (27 out of 1823):** This category reflects mismatches between the context and the intended target word, comprising 32.5%of all errors.
- **Sentence-level errors (3 out of 1823):** This category includes cases of poor phrasing or ambiguity, comprising 3.6% of all errors.

Overall, the verification process confirmed the overall quality of the benchmark, and flagged items were corrected accordingly.

3.4.3. Validate benchmark

After verification, we validated JAMAL using a stratified 25% sample (455 items) drawn proportionally from all 56 functional categories. Two native Arabic speakers with different academic backgrounds independently evaluated the items, ensuring a diverse perspective since commonsense benchmarks target shared everyday reasoning rather than specialized knowledge. All reviewers followed the instructions detailed in Appendix D, assessing each item across four criteria: **Cloze Predictability**, **Agreement with Reference**, **Category Validation**, and **Inferential Consistency**. The structured evaluation form ensured a systematic and consistent assessment of item quality.

- **Cloze predictability:** This criterion measures whether the reviewer's predicted word exactly matches the gold target word. The first reviewer achieved exact matches for 403 items (88.57%). Allowing for morphological variants sharing the same root increases this to 408 items (89.67%), and including valid synonyms further raises it to 417 items (91.65%). (In the evaluation of language models in later sections, we account for exact matches, root-based variants, and synonym matches.)

  The remaining 38 responses (8.35%) were distributed as follows: 1 item (0.22%) was left blank (target word: *knitting*), primarily due to reviewer oversight; 27 items (5.93%) were incorrect due to inattention; and 10 items (2.20%) reflected plausible alternative completions, indicating potential ambiguity in item design.

  For the second reviewer, exact matches were achieved for 409 items (89.89%). Allowing for root-related variants increases this to 411 items (90.33%), and including synonyms raises this to 415 items (91.21%). The remaining 40 responses (8.79%) were distributed as follows: 4 items (0.88%) were left blank; 30 items (6.59%) were incorrect due to inattention; and 6 items (1.32%) reflected plausible alternative completions.
- **Agreement with reference:** In this criterion, reviewers evaluated whether the expected word represented the most plausible continuation of the sentence. The expected word was considered correct even if the reviewer predicted a synonym or root variant (accounted for in the evaluation metrics), made a minor error, or left the item blank. An item was marked incorrect only when the reviewer identified a different plausible completion that was not synonymous with the target word.

  The first reviewer judged 445 items as correct (97.80%), while the second reviewer judged 449 items as correct (98.68%), resulting in an inter-annotator agreement of 96.7%.
- **Category validation:** Only one category misclassification was identified by the second reviewer. The word *embroidery*, in this context, was incorrectly classified under *Patterns and Design* instead of *Crafts and Hobbies*. The corresponding item was:

  "Maryam brought a piece of fabric and colored threads to decorate her old clothes. She decided to learn the art of ..." (*Crafts and Hobbies*).
- **Inferential consistency:** To test compliance with the CPARG format, both reviewers confirmed that the primary contextual cue is present in the first sentence and that successful inference requires the full context rather than only the second sentence. The target word is not explicitly mentioned, confirming that all items in the validation sample adhere to the CPARG format.

The validation results demonstrate strong inter-annotator agreement and support the semantic and categorical reliability of JAMAL.

3.4.4. Cross-cultural adaptation and migration pipeline

After the design, creation, and validation of JAMAL, we introduce a cross-cultural adaptation and migration pipeline to enable its extension to new cultural and linguistic settings. JAMAL is language-agnostic, as its structural organization across the functional, cognitive, and cultural axes is designed to generalize beyond Arabic. Adaptation to a new language or culture can therefore be achieved through three complementary strategies: translation of existing items, creation of new instances, and culture-specific generation of novel concepts, with all steps supported by human-in-the-loop validation.

Translation-based adaptation:

For cross-lingual transfer, automatic translation can be used to accelerate initial dataset construction. However, translation may alter syntactic structure and affect constraints such as the requirement that the target word appears in final position in the CPARG format. Therefore, all translated instances undergo post-verification to ensure compliance with the cloze structure and to preserve unambiguous target predictability.

Creation of new items:
When translation is insufficient or when extending JAMAL, new items are constructed directly following the CPARG format. This process consists of three steps:

1. Target word selection: Target words are selected for each category (e.g., food, music, clothing) using lexical and knowledge resources such as WordNet, ConceptNet, and BabelNet, as well as localized sources such as Wikidata and language model outputs. These resources provide culturally relevant candidate concepts.
2. Context construction: Two-sentence contexts are constructed around each target word. The first sentence introduces a contextual cue, while the second sentence is designed to end with the target word, which is removed during evaluation. This step can be automated using the pipeline described in Section 3.3.3, which generates CPARG-style contexts.
3. Human validation and refinement: Annotators verify that each instance satisfies CPARG structural constraints and is culturally appropriate for the target setting. When necessary, they refine wording or add minimal contextual cues to ensure naturalness and unambiguous target predictability.

Culture-specific adaptation:
The cultural axis of JAMAL requires additional care during adaptation, particularly for categories that contain culture-dependent knowledge. These include celebrations, films, arts and music. In such cases, adaptation is not strictly translational but involves functional cultural mapping, where concepts are replaced with culturally equivalent or culturally salient alternatives in the target setting.

For example, a celebration such as Eid may be mapped to Christmas or Diwali, a traditional dish to a locally equivalent food item, and a reference to a popular film to a culturally corresponding iconic movie. New culturally grounded concepts can also be introduced based on localized knowledge bases (e.g., Wikidata), native-speaker expertise, or culture-specific textual resources.

This strategy enables fine-grained adaptation not only across languages but also across dialects and regional varieties, supporting more localized and culturally sensitive evaluation settings.

4. Experiments and Discussion

We evaluated five Arabic language models on **JAMAL** using a Python script: AraGPT-base, MARBERT, AraBERT-large, FANAR-9B, and FANAR-27B.[3].

We evaluate model performance using four complementary metrics:

- **Exact match:** assigns a score of 1 if the predicted word exactly matches the gold target, and 0 otherwise.
- **Synonym match:** assigns a score of 1 if the predicted word is either identical to or a synonym of the target, based on Arabic WordNet, and 0 otherwise.
- **Same root:** uses the ISRI stemmer to determine whether the predicted and target words share the same morphological root.
- **Cosine similarity:** computes the semantic similarity between predicted and target words using FastText embeddings, yielding a score in $[0, 1]$.

Together, these metrics provide a multi-faceted evaluation of commonsense reasoning, capturing lexical accuracy, morphological similarity, and semantic relatedness.

4.1. Overall commonsense evaluation

As shown in Figure 3, same-root scores consistently exceed exact-match scores across all models, with AraGPT-base improving from 13.93% to 17.27%, MarBERT from 16.34% to 18.91%, AraBERT-large from 4.77% to 10.20%, FANAR-9B from 35.31% to 57.73%, and FANAR-27B from 62.88% to 72.53%. This pattern indicates that models frequently produce the correct lexical item with different conjugations rather than the exact target word.

Similarly, synonym-match scores exceed exact-match scores for all models: AraGPT-base rises from 13.93% to 17.87%, MarBERT from 16.34% to 19.30%, AraBERT-large from 4.77% to 10.86%, FANAR-9B from 35.31% to 58.77%, and FANAR-27B from 62.88% to 72.75%. This pattern is logical, as the synonym metric assigns a positive score when the model predicts either the exact word or one of its synonyms. Cosine similarity yields the highest scores overall, 37.32% for AraGPT-base, 42.99% for MarBERT, 26.67% for AraBERT-large, 61.85% for FANAR-9B, and 77.44% for FANAR-27B, suggesting that even when models do not output the expected word or a direct synonym, they often generate semantically related terms. These cases are analyzed further in subsection 4.2.

In model comparisons, the FANAR models achieve the highest performance across all metrics, with FANAR-27B substantially outperforming FANAR-9B, suggesting consistent benefits from scaling within the same architecture. Among smaller models, MARBERT (165M parameters) surpasses the larger AraBERT-large (370M parameters) across all metrics, and AraGPT-base (135M parameters) also outperforms AraBERT-large despite its smaller size.

These findings suggest that model size alone is not a reliable predictor of commonsense reasoning performance. Instead, differences in pretraining data, architectural design, and training objectives likely play an important role, although their individual contributions cannot be disentangled in our experiments.

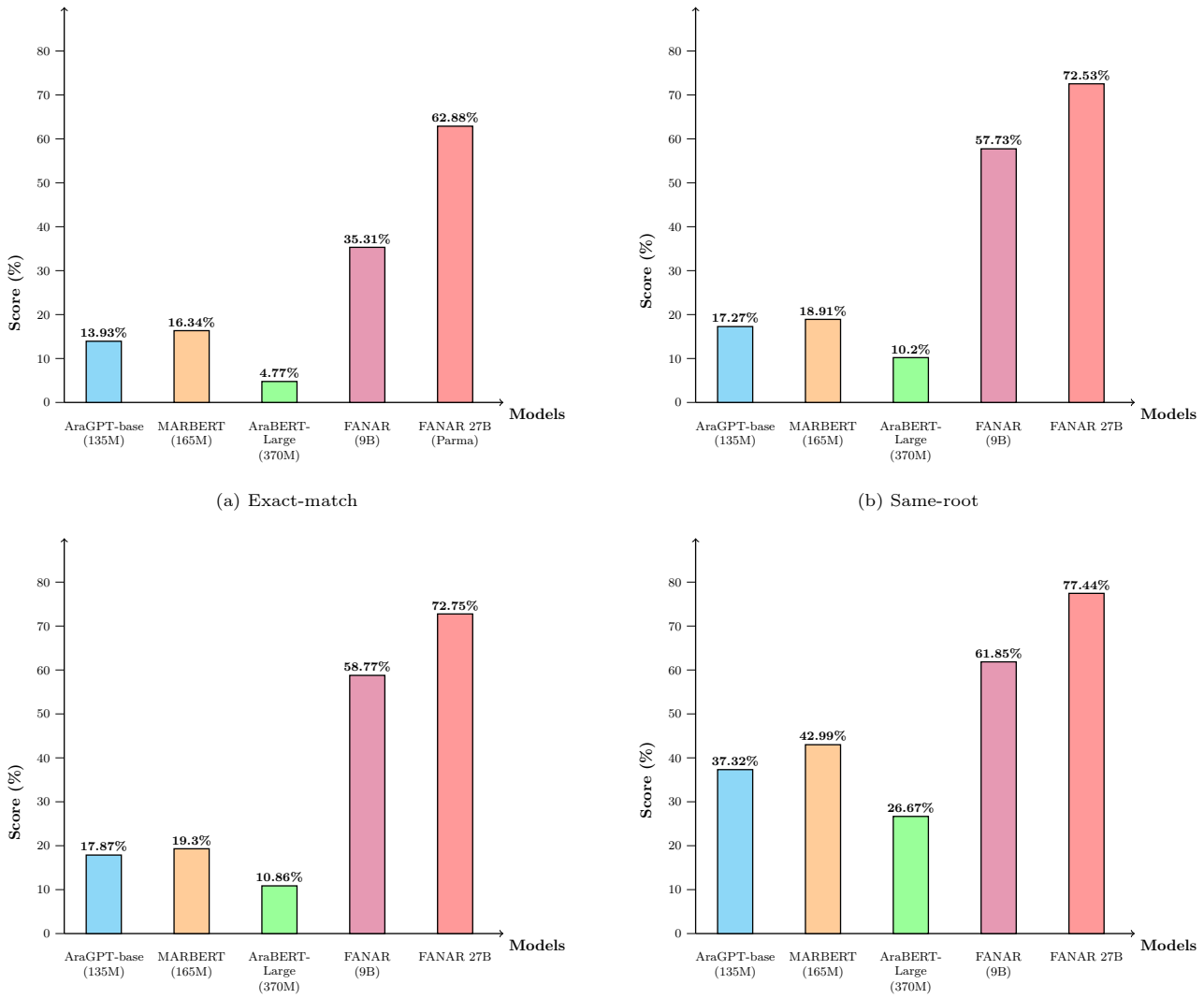
4.2. Error analysis

In this subsection, we present a qualitative analysis of model errors, followed by a theoretical interpretation of these behaviours grounded in prior research on transformer-based language models.

Error patterns:
We conducted a qualitative analysis of cases in which models received a score of zero under the Exact Match, Same Root, and Synonym Match metrics. Across AraGPT-base, MARBERT, AraBERT-large, FANAR-9B, and FANAR-27B, we identified several recurring error patterns. While not strictly mutually exclusive, these patterns capture distinct dimensions of model errors, ranging from semantic specificity and reasoning failures to cultural bias and syntactic interference.

- **Semantic neighbor substitution:** Models often predict semantically related concepts instead of the target word,

[3] The evaluation scripts and JAMAL are available at `https://github.com/BasmaSayah/An-ICF-guided-commonsense-benchmark`



(a) Exact-match

(b) Same-root

(c) Synonym-match

(d) Cosine similarity

Figure 3. Overall accuracy, same-root, synonym, and cosine similarity scores across evaluated models.

reflecting partial domain awareness. *Example:* For the context "When he traveled to Brazil, he saw a woman wearing a colorful dress moving her hands in a rapid rhythm. She was performing a [MASK]", the expected answer is *samba*. AraGPT-base predicts *the dancers*, MarBERT predicts *ballet*, and AraBERT-large predicts *hair*.

- **Hypernym substitution:** Specific items are replaced by broader categories, indicating a lack of precision. *Example:* "The melodies moved between different instruments in astonishing harmony. The audience listened attentively to the [MASK]." Expected: *symphony*. AraGPT-base predicts *song*, FANAR-27B predicts *playing*.
- **Hyponym confusion:** Models struggle to distinguish closely related hyponyms in specialized domains. *Example:* "She used a hooked needle and yarn to make a pillow cover. She learned the craft of [MASK]." Expected: *crochet*. FANAR-9B predicts *embroidery*, FANAR-27B predicts *tailoring*.
- **Causal overextension:** Models occasionally replace likely outcomes with extreme or abstract consequences. *Example:* "He walked on a thin rope above the circus. Everyone feared that he would [MASK]." Expected: *fall*. AraGPT-base predicts *die*.
- **Associative / Cultural bias:** Corpus-level or cultural associations can override contextual constraints. *Example:* "The boy listened to the song three times in a row. He was trying to memorize the [MASK]." Expected: *lyrics*. AraGPT-base, MarBERT, and FANAR-9B predict *the Quran*.
- **Context insensitivity / Syntactic priming:** Predictions may follow grammatical patterns while ignoring semantic fit. *Example:* "Ahmad raises his instrument to his chin and gets ready. He plays a sad song on the [MASK]." Expected: *violin*. AraGPT-base predicts *then*, FANAR-27B predicts *oud*.

These error patterns are consistent across all models, although they occur less frequently in FANAR-27B. Notably, high cosine similarity often reflects semantically related but

incorrect predictions rather than plausible completions, underscoring the limitations of embedding-based evaluation. Overall, the results highlight persistent weaknesses in lexical precision, causal reasoning, cultural awareness, and fine-grained semantic discrimination.

Interpretation of error patterns

To situate these findings within existing literature, we provide theoretical explanations grounded in prior work on transformer-based language models, highlighting interacting mechanisms involving semantic representation, data distribution effects, and syntactic , as well as decoding biases.

- **Semantic similarity errors (Semantic neighbor / Hypernym substitution)):** These errors arise when models select tokens that are semantically related to the target but differ in specificity. This includes both overgeneralized predictions (hypernyms) and contextual or associative neighbors. The behavior reflects the organization of semantic information in continuous embedding spaces, where related concepts are embedded in nearby regions [49, 50]. As a result, next-token prediction is influenced by competition among semantically similar candidates, often favoring higher-frequency or more generic alternatives with stronger prior probability [51].
- **Hyponym confusion:** This error type reflects difficulty in distinguishing closely related concepts within narrow or specialized domains. Models often capture the broader semantic field but fail to resolve fine-grained lexical distinctions between near-hyponyms. Prior work suggests that distributional representations compress fine-grained semantic differences into shared latent regions [6]. This effect is further amplified by the Zipfian distribution of language, where specific terms occur infrequently in training corpora [52]. Consequently, such terms receive fewer training signals, leading to weaker and less reliable representations [53].
- **Causal overextension:** This pattern reflects a tendency to generate salient or prototypical outcomes rather than contextually constrained causal consequences. Instead of explicitly modeling intermediate causal steps, models rely on high-probability continuations conditioned on surface context (i.e., shallow lexical and syntactic cues rather than deeper semantic or pragmatic understanding) [54]. This can result in overly extreme or exaggerated predictions when fine-grained causal constraints are not strongly represented in the training data.
- **Associative / Cultural bias errors:** These errors occur when strong corpus-level associations override local contextual constraints. Frequent co-occurrence patterns in pre-training data induce strong prior probabilities that may dominate contextual conditioning. As a result, culturally salient or high-frequency terms may be produced even when they are not contextually appropriate, reflecting well-documented bias propagation effects in large language models [55].
- **Context insensitivity / Syntactic priming:** These cases indicate reliance on surface-level syntactic or lexical patterns rather than deeper semantic integration. Models may produce locally fluent but globally inconsistent outputs when semantic constraints are weak or ambiguous. Prior studies show that transformer models often over-rely on local dependencies learned during training, at the expense of long-range semantic coherence [44].

Taken together, these theoretical perspectives suggest that the observed error patterns are not isolated failures but systematic behaviours arising from the interaction between embedding geometry, data distribution, and decoding biases. They also reflect a broader tension between reliance on high-probability training priors and the need for precise, contextually constrained reasoning.

4.3. Strengths and weaknesses of the models

Figures 4 and 5 show the root-match performance of AraGPT-base, MARBERT, AraBERT-large, FANAR-9B, and FANAR-27B across the 15 higher-level functional categories. Since the full set of 56 categories is too dense for visual presentation, we aggregate results at this higher level for clarity. The same figures also report performance across the three cognitive categories (everyday situations, general knowledge, and problem solving). Table 6 shows model performance across cultural scope categories (universal, mainstream, and Arabic).

4.3.1. Performance across functional and cognitive categories

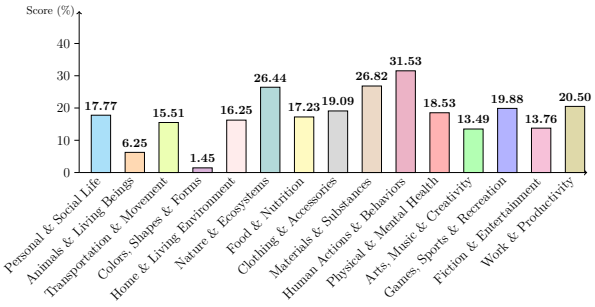
AraGPT-base demonstrates relatively balanced performance across the three commonsense branches (everyday situations, general knowledge, problem solving), suggesting uniform treatment of different reasoning types. At the functional level, it performs better in behavioural and environmentally grounded domains (e.g., Human Actions & behaviours, Nature & Ecosystems) but struggles in perceptual or visually grounded categories such as Colors and Shapes & Forms.

MarBERT achieves the strongest overall performance among the smaller models, particularly in human-centric and interaction-oriented domains (Personal & Social Life). This aligns with its pretraining on diverse social media text. However, like AraGPT-base, it underperforms in fine-grained perceptual categories. In the cognitive branches, MarBERT excels in Problem Solving compared to Everyday Situations and General Knowledge.

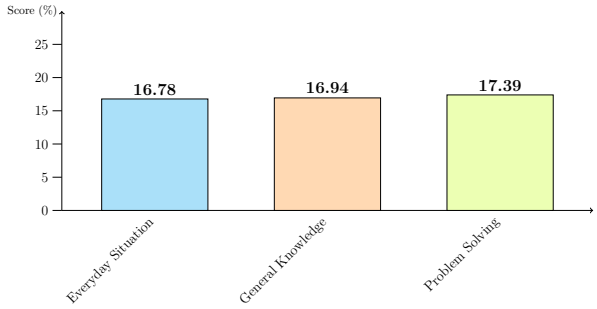
AraBERT-large shows lower performance across most categories despite being pretrained on the same dataset as AraGPT-base v2 [56]. AraBERT-large performs moderately in Materials & Substances, Games, Sports & Recreation, and Clothing & Accessories, and exhibits relatively better results in Problem Solving, suggesting that increased capacity may aid abstract reasoning.

Overall, these results confirm that training data, model architecture, pretraining objectives, and model size all influence commonsense reasoning. Models pretrained on socially rich and contextually diverse corpora, like MarBERT, exhibit stronger and more consistent performance, especially in human-centered domains.

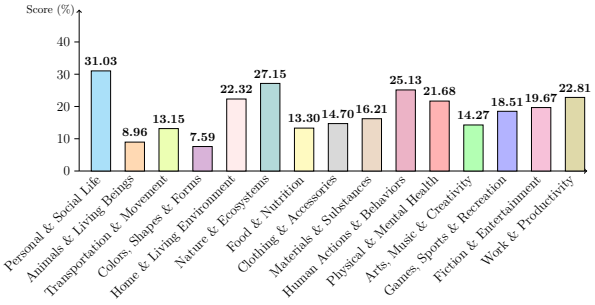
Figure 5 shows that FANAR-9B substantially outperforms smaller models across both cognitive and functional dimensions. It achieves balanced scores across the three cognitive categories: everyday situations, general knowledge, and problem solving; and demonstrates strong performance in human-centered and practical functional domains such as Food & Nutrition, Materials & Substances, Human Actions & behaviours, and Transportation & Movement. Perceptual and functional categories, such as Colors and Shapes & Forms, remain comparatively weaker, indicating room for improvement in fine-grained, concrete knowledge representation. Animals & Living Beings exhibits a comparable weakness.



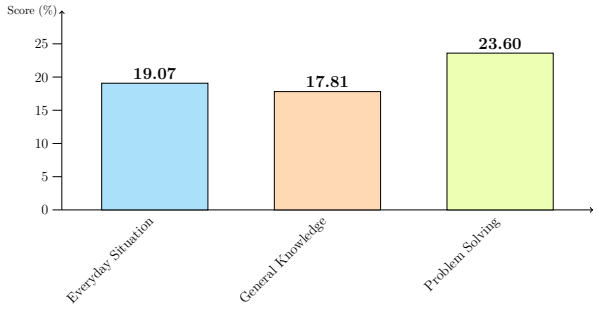
(A1) AraGPT-base root-match scores across life domains



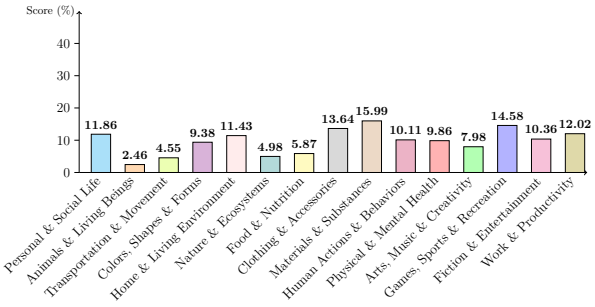
(A2) AraGPT-base root-match scores across cognitive dimension



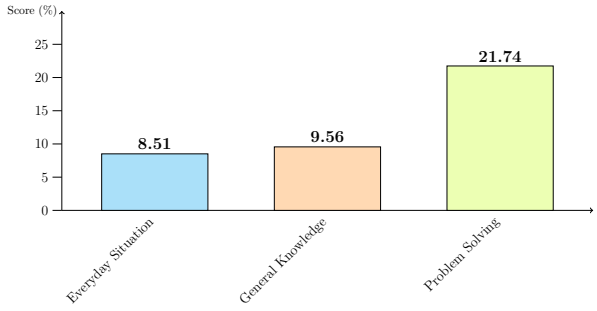
(B1) MarBERT root-match scores across life domains



(B2) MarBERT root-match scores across cognitive dimension



(C1) AraBERT-large root-match scores across life domains



(C2) AraBERT-large root-match scores across cognitive dimension

Figure 4. Strengths and weaknesses of AraGPT-base, MarBERT, and AraBERT-large.

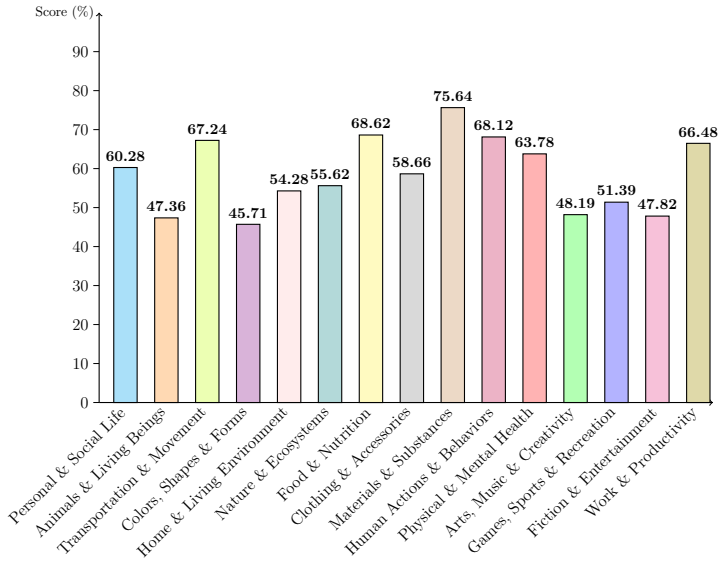
FANAR-27B consolidates the gains of FANAR-9B, outperforming it across nearly all functional and cognitive categories. Notable improvements are seen in visually grounded domains such as Colors, Shapes & Forms (+35%) , as well as in Animals & Living Beings (+21%), and Home & Living Environment (+23%). Cognitive categories also benefit, with Everyday Situations (+14%), General Knowledge (+15%), and Problem Solving (+11%) showing clear gains.

FANAR-27B achieves strong performance across functional categories and shows improved contextual integration compared to smaller models, with fewer extreme or irrelevant predictions. Its performance is closer to human cloze predictability in several visually grounded and everyday domains, suggesting improved semantic precision with increased model capacity. Despite these advances, the model still exhibits a tendency toward overgeneralization in some cases, and persistent gaps in problem solving indicate that challenges in higher-order reasoning remain.

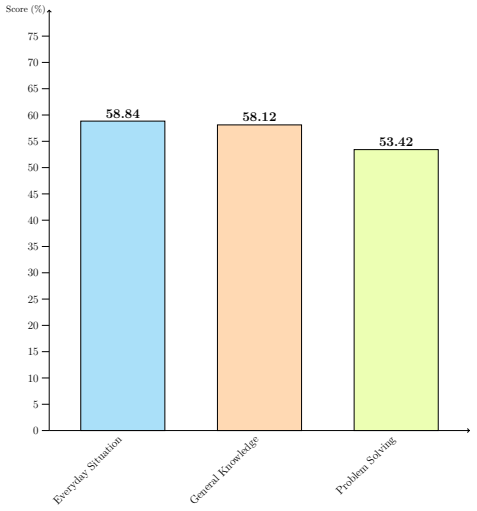
4.3.2. Performance analysis by cultural grounding

Building on the previous analysis, we examine model performance across the three cultural grounding categories: *Universal* (N = 1,615 items; knowledge shared across cultures, such as physical laws and daily routines), *Western/Global* (N = 78 items; mainstream concepts common in global media, such as karaoke or Monopoly), and *Arabic-specific* (N = 131 items; knowledge grounded in Arabic cultural contexts). This distribution reflects the natural skew in commonsense knowledge, where universal concepts are inherently more frequent than localized ones, thereby preserving realistic real-world frequencies rather than enforcing an artificially uniform design. The results, reported in Table 6 show that:

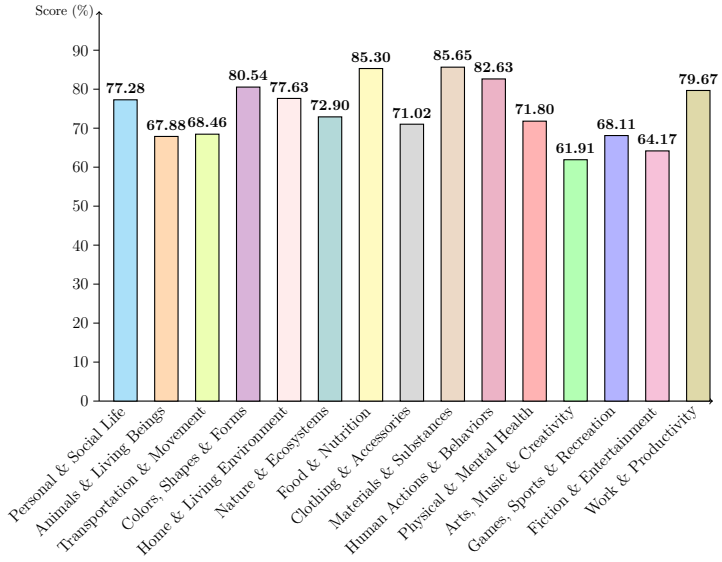
- Across all models, universal items consistently achieve higher performance than Arabic-specific items. For example, AraGPT-base achieves 14.49% on universal vs. 10.69% on Arabic, MarBERT 17.28% vs. 12.21%, AraBERT-large 4.95% vs. 2.29%, FANAR-9B 35.60% vs. 30.53%, and



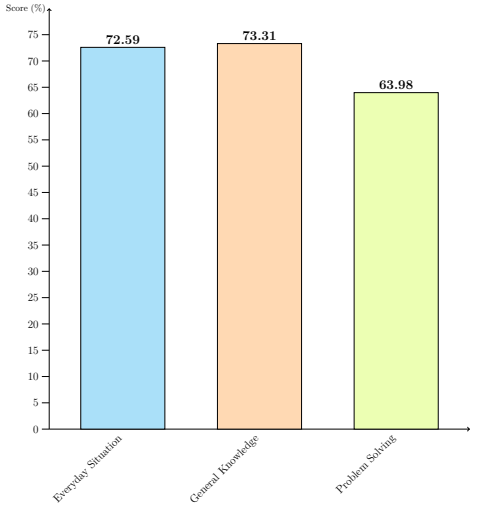
(A1) Fanar-9B root-match scores across life domains



(A2) Fanar-9B cognitive dimension scores



(B1) Fanar-27B root-match scores across life domains



(B2) Fanar-27B cognitive dimension scores

Figure 5. Strengths and weaknesses of Fanar-9B and Fanar-27B.

FANAR-27B 64.00% vs. 52.56%. This indicates a persistent performance gap between universal and culturally specific knowledge across all models.

- Performance on western/global items varies across models. AraGPT-base (7.69%) and MarBERT (3.85%) score lower on western items than on either Universal or Arabic-specific categories. In contrast, AraBERT-large (5.13%) and FANAR-9B (37.17%) achieve slightly higher scores on western items than on Universal items. FANAR-27B (54.96%) shows intermediate performance, with western scores lower than universal (64.00%) but higher than Arabic-specific items (52.56%). This variation may reflect differences in model capacity and exposure to culturally diverse content during training.
- Scaling from FANAR-9B to FANAR-27B yields higher absolute performance across all categories, with FANAR-27B reaching 64.00% (universal), 54.96% (western), and 52.56% (Arabic). However, the Universal–Arabic gap remains substantial, suggesting that while increased model capacity improves overall performance, disparities across cultural categories persist.

Overall, all models consistently show lower performance on Arabic-specific items than on universal items. This gap may reflect the comparatively limited representation of culturally specific knowledge in large-scale pretraining data rather than differences in item quality or ambiguity, particularly given the high human validation performance on these items.

Model	Overall	Universal (U)	Western (W)	Arabic (A)	Gap (U-A)	Gap (U-W)
AraGPT-base	13.93%	14.49%	7.69%	10.69%	3.80%	6.80%
MarBERT	16.34%	17.28%	3.85%	12.21%	5.06%	13.43%
AraBERT-large	4.77%	4.95%	5.13%	2.29%	2.66%	-0.17%
FANAR-9B	35.31%	35.60%	37.17%	30.53%	5.07%	-1.58%
FANAR-27B	62.88%	64.00%	54.96%	52.56%	9.06%	11.46%

Table 6. Performance comparison across cultural grounding categories. Overall accuracy is reported in the *Overall* column.

5. Conclusion

This study introduces **JAMAL**, a language-agnostic framework for evaluating commonsense reasoning in language models, alongside its Arabic instantiation as a benchmark. JAMAL adopts a three-axis taxonomy spanning a functional dimension (56 life-domain categories), a cognitive dimension (everyday situations, general knowledge, and problem-solving), and a cultural grounding axis (universal, western/global, and Arabic-specific knowledge). Together, these axes enable fine-grained, multi-dimensional evaluation that diagnoses model commonsense reasoning capabilities.

JAMAL is constructed through a multi-stage pipeline combining manual curation and LLM-assisted generation with human refinement. This hybrid design leverages the scalability of LLMs while maintaining quality through human verification. Its items follow the CPARG format [44], which imposes strict structural constraints and relies on subtle contextual cues to elicit commonsense inference.

The empirical evaluation of five Arabic language models reveals consistent performance gaps across functional, cognitive, and cultural dimensions, with FANAR-27B achieving the strongest overall results. Overall, JAMAL provides a structured and extensible benchmark for interpretable evaluation of commonsense reasoning, addressing a key gap in Arabic NLP and supporting future cross-lingual and culturally aware model assessment.

6. Limitations and Future Perspectives

While JAMAL provides a structured framework for evaluating contextual commonsense reasoning in Arabic language models, it is not exhaustive. Commonsense knowledge is broad, dynamic, and culturally situated. Future work could expand the benchmark with additional examples across functional life-domain categories and the problem-solving dimension, as well as include a wider range of culturally specific cases reflecting the diversity of Arabic-speaking communities.

The current benchmark primarily uses Modern Standard Arabic (MSA). Extending this work to Arabic dialects and more diverse linguistic settings is a natural direction for future research and would allow for a broader evaluation of model robustness across different forms of Arabic usage.

Although JAMAL has undergone manual verification, a full human performance baseline over the entire dataset remains an important direction for future work. Such a baseline would enable a more precise estimation of the human–model performance gap.

Finally, the CPARG format is a high-constraint cloze-based design in which each item presents a controlled two-sentence context with a single masked target. This reduces prompt variability and enables consistent assessment of contextual inference. However, it reflects a constrained form of language use rather than open-ended interaction, and thus captures only a subset of naturalistic language behavior.

Ethical Statement

No ethical approval was required for this study, as it did not involve human or animal subjects.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Declaration of competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability Statements

The data supporting the findings of this study are openly available in `https://github.com/BasmaSayah/An-ICF-guided-commonsense-benchmark`.

Credit authorship contribution statement

Basma Sayah: Conceptualization, Methodology, Data curation, Validation, Investigation, Writing – Original Draft. Attia Nehar: Methodology, Investigation, Writing – Review & Editing. Hadda Cherroun: Methodology, Investigation, Writing – Review & Editing. Slimane Bellaouar: Methodology, Investigation, Writing – Review & Editing. Firoj Alam: Methodology, Investigation, Resources, Writing – Review & Editing.

A. JAMAL Functional Dimension Categories Adapted from the International Classification of Functioning (ICF)

Benchmark Domain	Benchmark Subcategories	Mapped ICF Code and Component
Personal & Social Life	Family & Friends, Celebrations, Emotion, Communication	d7 "Interpersonal interactions and relationships"
Animals & Living Beings	Pets, Wildlife, Farming Animals, Insects & Small Creatures	e245 "Animals"
Transportation & Movement	Land Transport, Water Transport, Air Transport, Walking & Running & Movement	d4 "Mobility"
Colors, Shapes, and Forms	Basic Colors, Geometrical Shapes, Patterns & Designs	b156 "Perception of visual stimuli"; b160 "Thought functions"
Home & Living Environment	Buildings & Structures, Rooms & Spaces, Household Objects, Household Tools & Appliances	e1 "Products and technology"
Nature & Ecosystems	Plants & Trees, Bodies of Water, Weather & Seasons, Landscapes	e2 "Natural environment and human-made changes to environment"
Food & Nutrition	Types of Food, Cooking & Eating, Farming & Gathering	d550 "Eating"; d630 "Preparing meals"
Clothing & Personal Accessories	Fabrics & Textiles, Types of Clothes, Footwear, Accessories	d540 "Dressing"; e1 "Products and technology"
Materials & Substances	Natural Materials, Manufactured Materials, Chemical Substances	e1 "Products and technology"; e2 "Natural environment and human-made changes to environment"
Human Actions & Behaviors	Learning & Education, Work Actions, Play Actions, Helping & Caring	d1 "Learning and applying knowledge"; d8 "Major life areas"
Physical & Mental Health	Physical Abilities, Emotional Well-being, Illnesses & Conditions, Health Maintenance	b1 "Mental functions"; b7 "Neuromusculoskeletal and movement-related functions"
Arts, Music, and Creativity	Visual Arts, Performing Arts, Music, Crafts & Hobbies	d920 "Recreation and leisure"; b160 "Thought functions"
Games, Sports, and Recreation	Team Sports, Individual Sports, Board Games, Recreational Activities (e.g., hide and seek, tag, playground games)	d920 "Recreation and leisure"
Fiction & Entertainment	Stories & Books, Films & TV Shows, Fantasy & Myths	d920 "Recreation and leisure"; b160 "Thought functions"
Work & Productivity	Jobs & Occupations, Workplaces, Workplace Tools & Machinery, Business	d850 "Remunerative employment"; e1 "Products and technology"

Table 7. ICF-based functional dimension categories used in the JAMAL benchmark construction.

B. Distribution of JAMAL items across the 56 functional categories

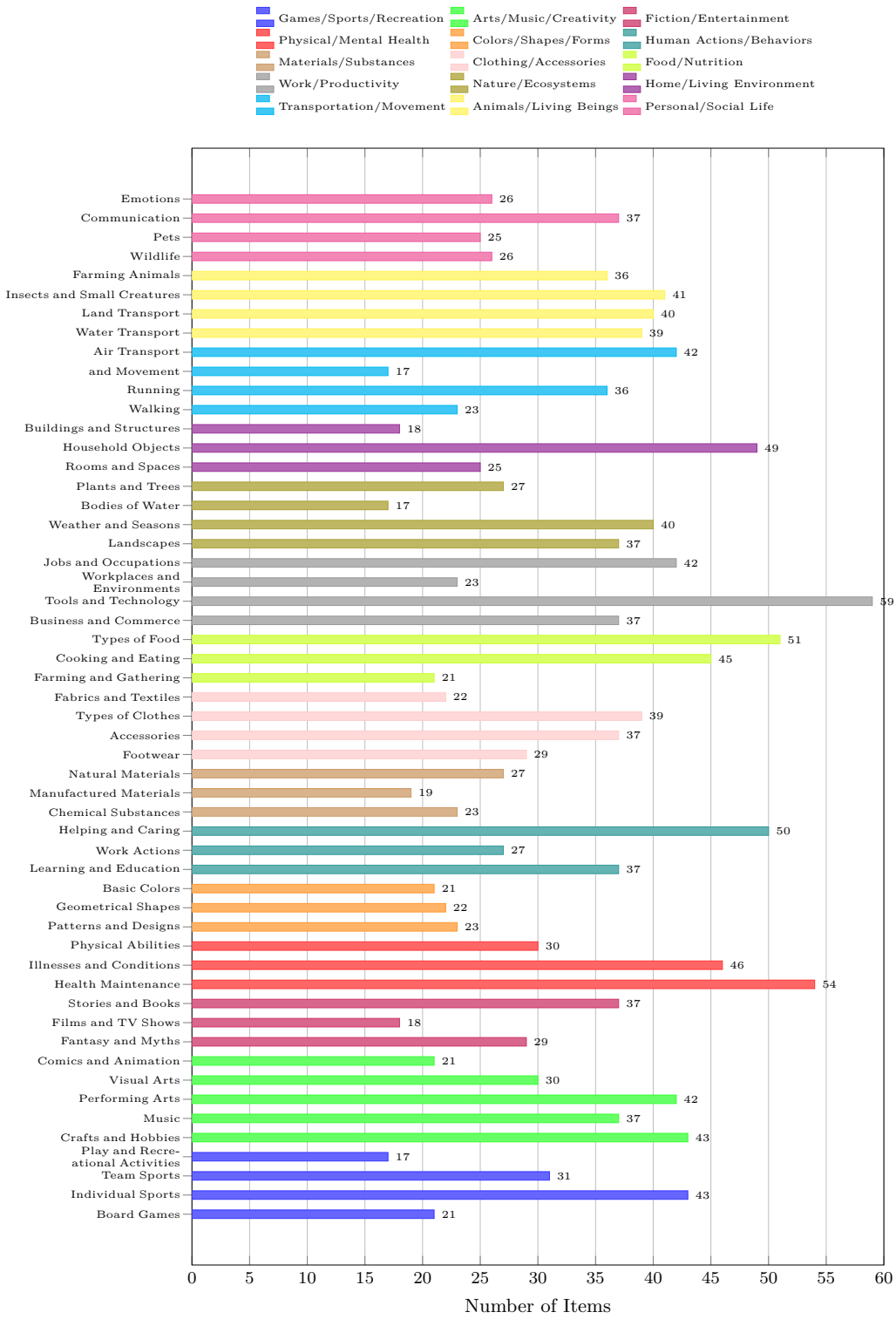


Figure 6. Distribution of JAMAL items across functional categories.

C. Benchmark Verification Guidelines

The reviewer was tasked with examining the entire benchmark. If a context failed to satisfy any of the following conditions, the reviewer was instructed to mark it with a ✗ in the designated box.

Verification Criteria

- Each context must consist of two sentences.
- The target word must be the most natural completion of the context.
- No alternative completion should be reasonably plausible upon reading the context.
- The target word should not be explicitly mentioned in the context; however, the first sentence should provide a clear hint toward it.
- The correct completion should not be inferable from the second sentence alone.
- The target word must be sufficiently clear and predictable so that a reader can infer it from the context.

Example of a Valid Context

✓ **Context:** Rosalyn announced happily, "Checkmate." She was going to become really good at playing. . .
Target word: chess
Justification: The first sentence provides a clear cue ("checkmate"), making "chess" the only plausible completion.

Examples of Invalid Contexts

1. **Context:** Ahmad went to the market. He bought some. . .
Target word: apples
✗ **Justification:** Multiple completions are possible (e.g., bread, milk, meat, fruit).
2. **Context:** Layla decided to bake fresh bread in the oven. She enjoyed the smell of. . .
Target word: bread
✗ **Justification:** The target word ("bread") is explicitly mentioned in the first sentence.
3. **Context:** The class time ended. The teacher started wiping the. . .
Target word: board
✗ **Justification:** The correct completion can be inferred from the second sentence alone.

D. Benchmark Validation Guidelines

The reviewers were tasked with completing sentence contexts and assessing multiple criteria for each benchmark item. The following guidelines outline the validation procedure.

Validation Objectives

- Assess *cloze predictability*: whether reviewer predictions match the target words.
- Assess *reference agreement*: whether the expected word represents the only natural completion.
- Determine *inferential consistency*: which sentence provides contextual clues for prediction.
- Verify *category alignment*: whether contexts properly belong to their assigned categories.

Validation Procedure

Step 1: Read the full context

Each item consists of two consecutive sentences. Read both sentences before making any judgments.

Step 2: Predict the final word

Write a single word that most naturally completes the context. Choose the most coherent completion based on linguistic intuition.

Step 3: Mark inference sources

✓ **Inferable from Sentence 1:** Mark if the first sentence contains the hint to the target word.
✓ **Inferable from Sentence 2:** Mark if the target word can be predicted from the second sentence alone.

Step 4: Validate category assignment

✓ **Matches Category 1:** Mark if the context clearly aligns with Category 1.
✓ **Matches Category 2:** Mark if the context clearly aligns with Category 2.

Step 5: Compare the predicted word to the expected word

✓ Mark **True** if the predicted word matches the expected word exactly.
✓ Mark **False** if the predicted word does not match the expected word.

Step 6: Assess the expected word

✓ Mark **True** if the expected word is the only plausible completion.
✓ Mark **False** if there exist other plausible completions that are not synonymous with the expected word.

Step 7: Document issues

Use the notes column to record:

- Unclear sentences.
- Mismatches between context and assigned categories.
- Difficulties in predicting the target word.
- Any other relevant observations.

Example of Proper Annotation

✓ **Context:** He put the letter in the envelope and dropped it in the mailbox. The worker noticed it was missing a . . .
Target word: stamp
Category 1 (Cognitive): Everyday Situation
Category 2 (Functional): Communication
Validator prediction: stamp
Annotation: Same word: ✓, Inferable from S1: ✓, Inferable from S2 alone: ✗, Matches Category 1: ✓, Matches Category 2: ✓, Uniqueness of the expected completion: ✓

General Instructions

- Rely on natural linguistic intuition; no external resources are permitted.
- Work independently without discussing items with the other reviewer.
- Leave cells blank when uncertain rather than guessing.
- Apply check marks only in designated columns.

Submission Requirements

Before final submission, reviewers must:

1. Ensure each prediction field contains exactly one word.
2. Verify check marks are placed only where appropriate.
3. Confirm all notes provide clear explanations where needed.
4. Save the completed file and send it to the researcher.

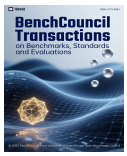
References

1. T. B. Brown *et al.*, "Language models are few-shot learners," 2020, doi: https://doi.org/10.48550/arXiv.2005.14165.
2. OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023, doi: https://doi.org/10.48550/arXiv.2303.08774. [Online]. Available: https://arxiv.org/abs/2303.08774
3. H. Touvron *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023. [Online]. Available: https://arxiv.org/abs/2307.09288
4. J. Bai *et al.*, "Qwen technical report," *arXiv preprint arXiv:2309.16609*, 2023, doi: https://doi.org/10.48550/arXiv.2309.16609.
5. G. Team *et al.*, "Gemini: A family of highly capable multimodal models," 2025. [Online]. Available: https://arxiv.org/abs/2312.11805
6. R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi, "HellaSwag: Can a machine really finish your sentence?" in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D. Traum, and L. Màrquez, Eds. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 4791–4800, doi: https://doi.org/10.18653/v1/P19-1472. [Online]. Available: https://aclanthology.org/P19-1472/
7. D. Hendrycks *et al.*, "Measuring massive multitask language understanding," 2021. [Online]. Available: https://arxiv.org/abs/2009.03300
8. G. Lai, Q. Xie, H. Liu, Y. Yang, and E. Hovy, "RACE: Large-scale ReAding comprehension dataset from examinations," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, M. Palmer, R. Hwa, and S. Riedel, Eds. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 785–794, doi: https://doi.org/10.18653/v1/D17-1082. [Online]. Available: https://aclanthology.org/D17-1082/
9. A. Abdelali *et al.*, "LAraBench: Benchmarking Arabic AI with large language models," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, Y. Graham and M. Purver, Eds. St. Julian's, Malta: Association for Computational Linguistics, Mar. 2024, pp. 487–520, doi: https://doi.org/10.18653/v1/2024.eacl-long.30. [Online]. Available: https://aclanthology.org/2024.eacl-long.30/
10. M. T. I. Khondaker, A. Waheed, E. M. B. Nagoudi, and M. Abdul-Mageed, "GPTAraEval: A comprehensive evaluation of ChatGPT on Arabic NLP," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 220–247, doi: https://doi.org/10.18653/v1/2023.emnlp-main.16. [Online]. Available: https://aclanthology.org/2023.emnlp-main.16/
11. F. Koto *et al.*, "ArabicMMLU: Assessing massive multitask language understanding in Arabic," in *Findings of the Association for Computational Linguistics: ACL 2024*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 5622–5640, doi: https://doi.org/10.18653/v1/2024.findings-acl.334. [Online]. Available: https://aclanthology.org/2024.findings-acl.334/
12. B. Mousi *et al.*, "AraDiCE: Benchmarks for dialectal and cultural capabilities in LLMs," in *Proceedings of the 31st International Conference on Computational Linguistics*, O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, and S. Schockaert, Eds. Abu Dhabi, UAE: Association for Computational Linguistics, Jan. 2025, pp. 4186–4218. [Online]. Available: https://aclanthology.org/2025.coling-main.283/
13. R. N. Almatham *et al.*, "BALSAM: A platform for benchmarking Arabic large language models," in *Proceedings of The Third Arabic Natural Language Processing Conference*, K. Darwish *et al.*, Eds. Suzhou, China: Association for Computational Linguistics, Nov. 2025, pp. 258–277, doi: https://doi.org/10.18653/v1/2025.arabicnlp-main.21. [Online]. Available: https://aclanthology.org/2025.arabicnlp-main.21/
14. S. Al-Khalifa, N. Durrani, H. Al-Khalifa, and F. Alam, "The landscape of arabic large language models," *Communications of the ACM*, vol. 68, no. 10, pp. 54–61, 2025, doi: https://doi.org/10.1145/3737453.
15. K. Smith *et al.*, "The origins of common sense in humans and machines," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 42, 2020, pp. 3–4. [Online]. Available: https://escholarship.org/uc/item/2367w9c4
16. S. K. Tawalbeh and M. Al-Smadi, "Is this sentence valid? an arabic dataset for commonsense validation," *CoRR*, vol. abs/2008.10873, 2020. [Online]. Available: https://arxiv.org/abs/2008.10873
17. N. Sengupta *et al.*, "Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models," 2023. [Online]. Available: https://arxiv.org/abs/2308.16149
18. M. S. Bari *et al.*, "ALLam: Large language models for arabic and english," in *The Thirteenth International Conference on Learning Representations*, 2025. [Online]. Available: https://openreview.net/forum?id=MscdsFVZrN
19. F. Team *et al.*, "Fanar: An arabic-centric multimodal generative ai platform," 2025. [Online]. Available: https://arxiv.org/abs/2501.13944
20. World Health Organization, *International Classification of Functioning, Disability and Health (ICF)*. Geneva: World Health Organization, 2001. [Online]. Available: https://www.who.int/classifications/icf/en/
21. R. Zellers, Y. Bisk, R. Schwartz, and Y. Choi, "SWAG: A large-scale adversarial dataset for grounded commonsense inference," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, Eds. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 93–104, doi: https://doi.org/10.18653/v1/D18-1009. [Online]. Available: https://aclanthology.org/D18-1009/
22. A. Talmor, J. Herzig, N. Lourie, and J. Berant, "CommonsenseQA: A question answering challenge targeting commonsense knowledge," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein,

C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4149–4158, doi: https://doi.org/10.18653/v1/N19-1421. [Online]. Available: https://aclanthology.org/N19-1421/

23. M. Sap, H. Rashkin, D. Chen, R. Le Bras, and Y. Choi, "Social IQa: Commonsense reasoning about social interactions," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 4463–4473, doi: https://doi.org/10.18653/v1/D19-1454. [Online]. Available: https://aclanthology.org/D19-1454/
24. Y. Bisk, R. Zellers, R. Le Bras, J. Gao, and Y. Choi, "Piqa: Reasoning about physical commonsense in natural language," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 7432–7439, doi: https://doi.org/10.1609/aaai.v34i05.6239. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/6239
25. M. Chen, M. D'Arcy, A. Liu, J. Fernandez, and D. Downey, "CODAH: An adversarially-authored question answering dataset for common sense," in *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, A. Rogers, A. Drozd, A. Rumshisky, and Y. Goldberg, Eds. Minneapolis, USA: Association for Computational Linguistics, Jun. 2019, pp. 63–69, doi: https://doi.org/10.18653/v1/W19-2008. [Online]. Available: https://aclanthology.org/W19-2008/
26. B. Y. Lin, S. Lee, X. Qiao, and X. Ren, "Common sense beyond English: Evaluating and improving multilingual language models for commonsense reasoning," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Online: Association for Computational Linguistics, Aug. 2021, pp. 1274–1287, doi: https://doi.org/10.18653/v1/2021.acl-long.102. [Online]. Available: https://aclanthology.org/2021.acl-long.102/
27. H. Santos, A. M. Mulvehill, K. Shen, M. Kejriwal, and D. L. McGuinness, "Tg-csr: A human-labeled dataset grounded in nine formal commonsense categories," *Data in Brief*, vol. 51, p. 109666, 2023, doi: https://doi.org/10.1016/j.dib.2023.109666. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352340923007515
28. S. Saha, P. Yadav, L. Bauer, and M. Bansal, "ExplaGraphs: An explanation graph generation task for structured commonsense reasoning," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 7716–7740, doi: https://doi.org/10.18653/v1/2021.emnlp-main.609. [Online]. Available: https://aclanthology.org/2021.emnlp-main.609/
29. W. Zhan *et al.*, "Score: Benchmarking long-chain reasoning in commonsense scenarios," 2025. [Online]. Available: https://arxiv.org/abs/2503.06218
30. H. Mozannar, E. Maamary, K. El Hajal, and H. Hajj, "Neural Arabic question answering," in *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, W. El-Hajj *et al.*, Eds. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 108–118, doi: https://doi.org/10.18653/v1/W19-4612. [Online]. Available: https://aclanthology.org/W19-4612/
31. J. L. Lee *et al.*, "Massively multilingual pronunciation modeling with WikiPron," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, N. Calzolari *et al.*, Eds. Marseille, France: European Language Resources Association, May 2020, pp. 4223–4228. [Online]. Available: https://aclanthology.org/2020.lrec-1.521/
32. S. Lamsiyah *et al.*, "Arabicsense: A benchmark for evaluating commonsense reasoning in arabic with large language models," in *Proceedings of the 4th Workshop on Arabic Corpus Linguistics (WACL-4)*, 2025, pp. 1–11. [Online]. Available: https://aclanthology.org/2025.wacl-1.1/
33. A. Sadallah *et al.*, "Commonsense reasoning in arab culture," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025, pp. 7695–7710, doi: https://doi.org/10.18650/ACL.2025.380.
34. A. Hornby and J. Turnbull, *Oxford Advanced Learner's Dictionary of Current English*. Oxford University Press, 2015.
35. F. Ilievski, A. Oltramari, K. Ma, B. Zhang, D. L. McGuinness, and P. Szekely, "Dimensions of commonsense knowledge," *Knowledge-Based Systems*, vol. 229, p. 107347, 2021, doi: https://doi.org/10.1016/j.knosys.2021.107347. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705121006092
36. D. B. Lenat, "Cyc: a large-scale investment in knowledge infrastructure," *Commun. ACM*, vol. 38, no. 11, p. 33–38, Nov. 1995, doi: https://doi.org/10.1145/219717.219745.
37. M. E. Whiting and D. J. Watts, "A framework for quantifying individual and collective common sense," *Proceedings of the National Academy of Sciences*, vol. 121, no. 4, p. e2309535121, 2024, doi: https://doi.org/10.1073/pnas.2309535121.
38. H. A. Simon and A. Newell, *Human problem solving: The state of the theory in 1970*. American Psychological Association, 1971, vol. 26, no. 2, doi: https://doi.org/10.1037/h0030806.
39. R. C. Schank and R. P. Abelson, *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Psychology press, 2013.
40. P. Schuyler, *Common Sense*. Los Angeles, CA: Higher Shelf Publishing, 2003. [Online]. Available: https://www.amazon.com/Common-Sense-Peter-Schuyler/dp/1932636021
41. R. Pesonen, "Casual reasoning : A social ecological look at human cognition and common sense," 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:202257933
42. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *North American Chapter of the Association for Computational Linguistics*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:52967399
43. A. Radford, "Improving language understanding with unsupervised learning," *OpenAI Res*, 2018. [Online]. Available: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

44. A. Ettinger, "What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 34–48, 2020, doi: https://doi.org/10.1162/tacl_a_00298. [Online]. Available: https://aclanthology.org/2020.tacl-1.3/
45. K. D. Federmeier and M. Kutas, "A rose by any other name: Long-term memory structure and sentence processing," *Journal of Memory and Language*, vol. 41, no. 4, pp. 469–495, 1999, doi: https://doi.org/10.1006/jmla.1999.2660. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0749596X99926608
46. K. Sakaguchi, R. L. Bras, C. Bhagavatula, and Y. Choi, "Winogrande: an adversarial winograd schema challenge at scale," *Commun. ACM*, vol. 64, no. 9, p. 99–106, Aug. 2021, doi: https://doi.org/10.1145/3474381.
47. R. H. Ennis, W. L. Gardiner, R. Morrow, D. Paulus, and L. Ringel, *The cornell class-reasoning test, Form X*, 1964.
48. S. Ouyang, J. M. Zhang, M. Harman, and M. Wang, "An empirical study of the non-determinism of chatgpt in code generation," *ACM Trans. Softw. Eng. Methodol.*, vol. 34, no. 2, Jan. 2025, doi: https://doi.org/10.1145/3697010.
49. K. Ethayarajh, "How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 55–65, doi: https://doi.org/10.18653/v1/D19-1006. [Online]. Available: https://aclanthology.org/D19-1006/
50. W. Gurnee, N. Nanda, M. Pauly, K. Harvey, D. Troitskii, and D. Bertsimas, "Finding neurons in a haystack: Case studies with sparse probing," *Transactions on Machine Learning Research*, 2023. [Online]. Available: https://openreview.net/forum?id=JYs1R9IMJr
51. A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, "The curious case of neural text degeneration," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=rygGQyrFvH
52. G. K. Zipf, *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Cambridge, MA: Addison-Wesley, 1949.
53. O. Levy, Y. Goldberg, and I. Dagan, "Improving distributional similarity with lessons learned from word embeddings," *Transactions of the Association for Computational Linguistics*, vol. 3, pp. 211–225, 2015, doi: https://doi.org/10.1162/tacl_a_00134. [Online]. Available: https://aclanthology.org/Q15-1016/
54. A. Srivastava *et al.*, "Beyond the imitation game: Quantifying and extrapolating the capabilities of language models," *Transactions on Machine Learning Research*, 2023, featured Certification. [Online]. Available: https://openreview.net/forum?id=uyTL5Bvosj
55. E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big? ," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 610–623, doi: https://doi.org/10.1145/3442188.3445922.
56. W. Antoun, F. Baly, and H. Hajj, "AraGPT2: Pretrained transformer for Arabic language generation," in *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, N. Habash *et al.*, Eds. Kyiv, Ukraine (Virtual): Association for Computational Linguistics, Apr. 2021, pp. 196–207. [Online]. Available: https://aclanthology.org/2021.wanlp-1.21/



BenchCouncil Transactions on Benchmarks, Standards and Evaluations, 2026

DOI: https://doi.org/10.66834/798kjf21
Research Article

RESEARCH ARTICLE

Design and Evaluation of an Interpretable Multimodal Deep Learning Framework for Early Alzheimer's Disease Detection

Shehu Mohammed,[1,*] Neha Malhotra[1] and Anmol Singh Rai[2]

[1]School of Computer Applications, Lovely Professional University, Phagwara, India and [2]Shrimann Superspeciality Hospitals, Jalandhar, India
[*]Corresponding author. mohammedshehumafara@gmail.com
Received on 22 December 2025; Accepted on 24 June 2026

Abstract

Alzheimer's disease is a progressive neurodegenerative disorder that significantly impairs memory and cognitive functions and affects over 55 million people worldwide. The successful management and planning require early and accurate diagnosis. Conventional radiological assessment is often subjective and time-consuming, which highlights the need for automated and reliable diagnostic solutions. Most deep learning models show promise for classifying neuroimaging data, but they tend to be less computationally efficient and less interpretable, and they cannot be integrated into patient-centric processes. The gap between developing diagnostic algorithms with high accuracy and implementing them in a supportive framework that includes patients and caregivers is large. This paper introduces a comprehensive, hybrid framework that addresses these gaps. We present a dual-modality diagnostic system: a deep learning pipeline using EfficientNetV2-S for CT scan classification, complemented by a Feedforward Neural Network (FNN) that analyses structured clinical data for holistic patient assessment. This diagnostic core is integrated into a user-friendly graphical user interface (GUI) and supplemented by "NeuroBot," an AI-powered chatbot that provides domain-specific information and support. The two models have been trained using the transfer learning method on a curated dataset of 30,000 brain CT slices. The EfficientNetV2-S model achieved an accuracy of 98.19%. After hyperparameter tuning, the FNN model achieved an optimised accuracy of 87.21%. The importance of the features addressed by the models was proved with the help of the statistical t-tests of the corresponding clinical data. The integrated system enables a scalable, translatable, and patient-centered system to improve the early analysis and treatment of Alzheimer's disease.

Key words: Alzheimer's Disease, CT Scan Classification, EfficientNetV2-S, Deep Learning, AI Chatbot, Medical Imaging, Transfer Learning

1. Introduction

Alzheimer's disease (AD) is the most prevalent form of dementia and primarily affects the aging population, accounting for approximately 60-70% of dementia cases worldwide. It is a neurodegenerative, advanced, and progressive disorder that slowly affects the cognitive functions, starting with the memory and then with the dysfunction of judgment, language, and behaviour. The disease has a significant social impact, placing substantial emotional and financial burdens on patients, families, and healthcare systems worldwide. Early and accurate diagnosis is therefore critical important. The timely diagnosis will enable patients and the individuals who provide care to investigate the treatment interventions, develop useful long-term care plans, and eventually improve the quality of life of patients as the disease advances[1, 2].

The neuroimaging procedures are the core of the AD diagnostics process, and they allow clinicians to observe structural changes of the brain typical of the disease, e.g., cortical atrophy and ventricular enlargement. Among these techniques, Computed Tomography (CT) scans serve as a vital tool. CT imaging is particularly valuable due to its rapid acquisition time, widespread availability, and relative cost-effectiveness compared to other modalities like Magnetic Resonance Imaging (MRI), making it a cornerstone of initial neurological workups, especially in resource-limited settings[3]. However, the conventional analysis of these scans is not without its challenges. The diagnostic process is a subjective one and depends on the radiologist's interpretation, and this cannot be adequately performed in a short time, besides being prone to inter-rater inconsistency and human error, especially in detecting the changes at the initial stages of AD.

These diagnostic limitations have been a primary driver for the development of computer-aided diagnostic (CAD) systems, a field that has been revolutionized by the advent of deep learning. As extensively documented in systematic reviews, both machine learning and deep learning techniques have demonstrated profound success in the automated analysis of medical images[4]. CNNs have been exceptionally skilled at this task, in particular. Their architecture allows them to automatically learn and identify complex, hierarchical patterns within visual data, enabling the detection of subtle pathological indicators in MRI and brain CT slices that may be imperceptible to the human eye[5, 6]. This capability has spurred the creation of numerous advanced models, including various hybridized deep learning approaches, all aimed at pushing the boundaries of diagnostic accuracy and reliability[7].

Despite this remarkable progress, a significant gap persists between the development of high-accuracy algorithms in a research setting and their practical deployment in clinical workflows. Many state-of-the-art CNN models face considerable hurdles, including the need for substantial computational resources, which limits their real-time application. Moreover, the inherent "black-box" nature of many deep learning models can be a barrier to clinical adoption; for a diagnosis to be trusted, clinicians must be able to understand and verify the reasoning behind it, a critical factor in differential diagnosis[8]. Beyond the technical challenges, most existing research has focused almost exclusively on the diagnostic algorithm itself. This narrow focus often neglects the broader clinical ecosystem and the critical need for an integrated, user-friendly system that not only provides a diagnosis but also supports patients and caregivers with accessible, contextual information[9].

To overcome these complex issues, this paper proposes a new and unique hybrid framework that combines the high-performance diagnostic engine with an interactive patient support system that is based and comprehensive. Our primary contribution is the development of a dual-model deep learning pipeline that leverages a powerful and efficient EfficientNetV2-S architecture for the classification of brain CT slices with exceptional accuracy and computational efficiency[10]. The core novelty of our work lies in embedding these performant and interpretable models, which utilize Grad-CAM for visual explanations, within a complete, end-to-end ecosystem. This system includes an intuitive graphical user interface (GUI) for seamless interaction by clinicians and an AI-powered chatbot, NeuroBot, designed to answer AD-related queries from patients and caregivers. By moving beyond mere classification, this holistic approach creates a practical, scalable, and supportive tool designed to enhance early Alzheimer's detection and improve the overall standard of patient care[11]. Furthermore, to capture non-visual risk factors and cognitive metrics that are crucial for a comprehensive diagnosis, our framework integrates a secondary pathway that leverages a robust FNN to analyze structured patient clinical records.

Although many deep learning models have achieved high diagnostic accuracy for Alzheimer's disease detection, most studies focus primarily on algorithm development and evaluation using experimental datasets. These models are rarely integrated into user-friendly systems that support clinical workflows, patient interaction, or caregiver guidance. As a result, there remains a gap between high-performance diagnostic algorithms and practical systems that can be deployed in real clinical environments. The proposed framework addresses this gap by integrating the diagnostic models within an interactive ecosystem that includes a graphical user interface and an AI-based assistant to support clinicians, patients, and caregivers.

The main contributions of this study can be listed as follows:

- Development of a dual-modality deep learning framework, which uses EfficientNetV2-S for CT image classification, and a Feedforward Neural Network for structured clinical data analysis.
- The application of explainable AI, which uses Grad-CAM for region highlighting on brain CT slices, thereby improving their interpretability.
- Development of a user-centric deployment platform, which includes a GUI and an AI-based chatbot, referred to as NeuroBot.
- The validation of clinical features through statistical analysis, which includes independent sample t-tests and correlation analysis to ascertain the significance of cognitive and lifestyle factors, such as MMSE and ADLs.

2. Literature Review

The ML and DL applied to the neurology practice have completely changed the Alzheimer Disease (AD) diagnostic process and provided meaningful and data-intuitive answers to the current practices. The evolution of these techniques has been rapid, moving from foundational models to highly sophisticated, specialized architectures. Early research in automated AD detection was primarily centered on classical machine learning algorithms. Models such as Support Vector Machines (SVMs), Random Forests, and Decision Trees were applied to neuroimaging data, but their efficacy was often constrained by a reliance on handcrafted feature extraction[12]. This process, which required extensive domain knowledge to manually define and extract relevant features like hippocampal volume or cortical thickness, was not only labor-intensive but also limited the models' ability to discover novel, complex patterns within the data. Despite the need to take such measures, the subsequent emergence of deep learning, and, particularly, the Convolutional Neural Networks (CNNs) became a game-changer in the matter. The CNNs have revolutionised the study of medical images; they enable end-to-end learning of hierarchical features beneath the raw pixel data. This capability has led to a proliferation of studies demonstrating the accurate prediction and diagnosis of AD using a variety of deep learning models, which consistently outperform their traditional ML counterparts[13, 14].

The current research in the area, however, has predominantly been interested in the application of deep learning to analyze neuroimaging images, and, more specifically, in Magnetic Resonance Imaging (MRI) and, more recently, in Computed Tomography (CT) scans. The authors have experimentally shown that applying advanced image processing and enhancement techniques before model training improves AD detection accuracy[15]. Recognizing that a single data source may not capture the complexity of AD pathology, numerous studies have investigated multimodal deep learning methods. These models integrate data from different neuroimaging modalities (e.g., structural MRI, functional MRI, and PET scans) to create a more comprehensive and robust diagnostic picture[16]. The sheer volume of research has produced a rich body of literature that systematically reviews the diverse array of AD detection techniques, highlighting the consistent and rapid progress in diagnostic accuracy and model sophistication[17]. Furthermore, the application of these powerful models has expanded beyond

AD to include the diagnosis of related neurodegenerative conditions, such as tauopathies, thereby demonstrating their broader versatility and clinical potential[18].

Since the field has become mature, the quest to achieve increasingly high accuracy has given rise to the emergence of new and complicated model architectures. Hybrid systems and ensemble-based systems are emerging to the fore. These advanced approaches combine the predictive strengths of multiple deep learning models to create more robust, reliable, and generalizable identification systems that are less prone to the biases of a single model[19, 20]. While neuroimaging remains the primary data source, some innovative models have been designed to diagnose AD based exclusively on structured patient clinical records, providing a valuable alternative or complementary diagnostic pathway that does not require imaging[21]. This has contributed to the development of sophisticated architectures capable not only of binary classification but also of detecting and differentiating between the various stages of AD, from early-stage Mild Cognitive Impairment (MCI) to advanced dementia, which is crucial for tailoring patient care[22, 23].

For these powerful deep learning models to transition from research laboratories to real-world clinical environments, two practical factors have become paramount: computational efficiency and model interpretability. Clinically viable diagnostic tools have been developed to continue to focus on the transfer learning application. The fine-tuning of large models, which have been trained on general image datasets, on smaller medical imaging datasets is a very effective way to achieve state-of-the-art performance with small datasets[24]. This strategy has been a central theme in the broader investigation of deep learning for enhancing early detection and supporting clinical decision-making[25, 26]. Consequently, a significant portion of modern research is focused on developing models for early detection, as this is the stage at which interventions are most likely to be effective[27]. To further improve performance, some researchers have proposed integrative models that combine the strengths of traditional ML with advanced deep learning architectures[28]. Finally, to address the "black-box" problem, there is a growing emphasis on creating explainable AI (XAI). The development of specialized, attention-based explainable networks and custom models, such as ADD-Net, underscores the field's commitment to creating diagnostic tools that are not only accurate and efficient but also transparent and trustworthy for clinical use[29, 30].

Although many researchers have reported positive results in detecting Alzheimer's disease using deep learning models in their literature, a wide range of methodological variations can be seen in their approaches. Most of the existing literature has focused on MRI-based deep learning models using CNN-based architectures and has achieved good accuracy using ADNI databases, ranging from 85% to 96%. However, a few researchers have also used multimodal data for the accurate prediction of Alzheimer's disease using MRI and PET scans. However, these techniques require expensive hardware. In addition, fewer researchers have focused on CT-based deep learning models for detecting Alzheimer's disease. Moreover, few researchers have focused only on prediction models without using interpretability mechanisms or deployment frameworks. Although various researchers have used various models, such as ADD-Net and attention-based models, to explain their models, these models are not used in deployment frameworks. However, in the proposed model, a CT-based deep learning model was used in conjunction with a structured clinical data model, an explainable AI model using a Grad-CAM mechanism, and a user-centric deployment model using a GUI and an AI chatbot support system[31].

3. Methodology and Experiment

The proposed framework integrates a dual-model deep learning pipeline for AD classification with an interactive user interface and an AI chatbot. The architecture is designed for accuracy, interpretability, and user engagement.

3.1. Dataset and System Architecture

The general procedure of our system is illustrated in Figure 1. The training and evaluation dataset consists of 10,240 brain CT slices, evenly distributed between Alzheimer's and non-demented cases. In addition, a clinical dataset containing 2,149 patient records was used for structured data analysis. Figure 2 presents the distribution of diagnoses in the clinical dataset, with 64.6% classified as Non-Demented and 35.4% as Demented. All the images were resized to an average of 224×224 pixels and normalised. Random rotations, horizontal flips, and zooming were used as data augmentation techniques to improve the model's robustness.

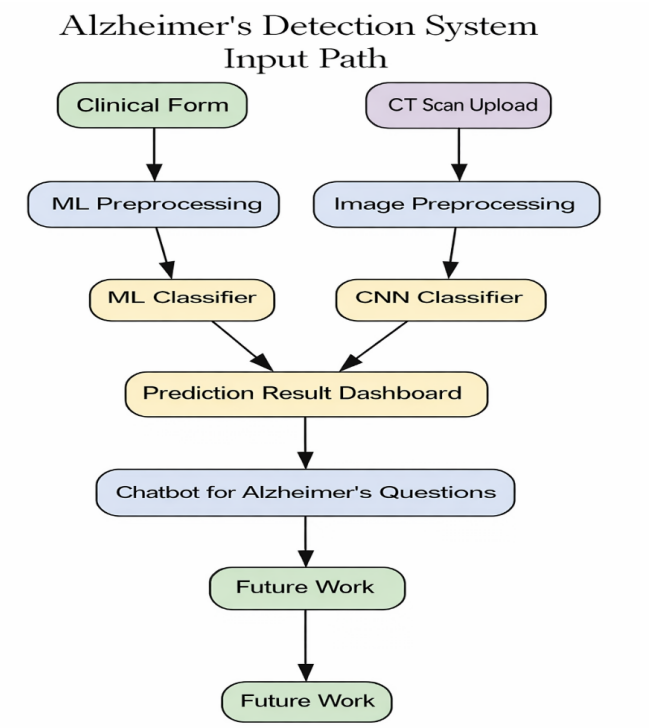


Figure 1. Alzheimer's Detection System Architecture.

A stringent, automated preprocessing pipeline was applied to the clinical data to prepare it for the FNN model. The first split of the features was into numerical (e.g., Age, MMSE) and categorical ones (e.g., Gender, Smoking status). One-hot encoding was then applied to all categorical features in order to put them in numerical form, and all numerical features were rescaled to a normalized range. This ensures that all features of the corresponding role in the prediction of the model are not affected by different scales. All this transformation process was stored as a single object of pipeline in order to make sure that

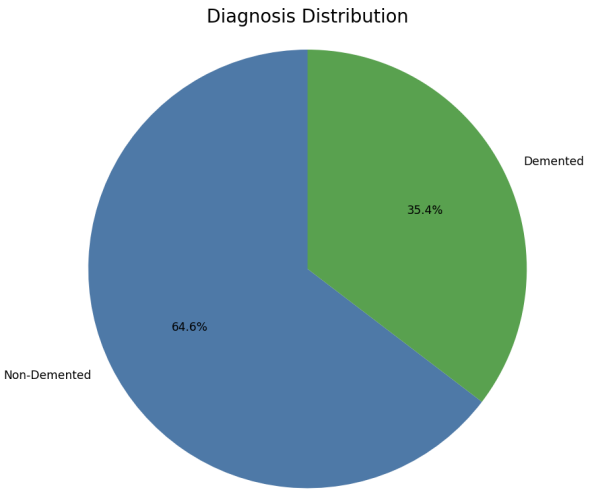


Figure 2. Diagnosis Distribution in the Clinical Dataset.

whenever new information is keyed in the prediction process, it would be processed similarly to the training information.

The CT imaging dataset used in this study consisted of approximately 30,000 brain CT slices derived from the Shrimaan Super Specialty Hospital repository and clinical datasets for Alzheimer's disease research. Of these, 10,240 CT slices were selected for training and evaluation purposes. The images were divided into training, validation, and testing subsets using a 70–10–20 split. In addition, an extended testing pool was used during the final evaluation, resulting in approximately 4,500 CT slices used for performance assessment. Each CT slice in the dataset was linked to a subject-level diagnosis label: Demented or Non-Demented.

To avoid data leakage in the model, the clinical dataset was split at the patient level instead of the slice level. This implies that brain CT slices for a specific patient were only included in either the training or testing dataset. Finally, the clinical dataset was split into 70% for training, 10% for validation, and 20% for testing.

The clinical dataset included 2,149 patient records, which included various physiological measures, cognitive measures such as Mini-Mental State Examination (MMSE), and lifestyle factors. For handling missing values in the clinical dataset, median imputation was applied for numerical features and most frequent imputation for categorical features. This particular process guaranteed the proper transformation of both training and testing data.

3.2. System Deployment Architecture

The framework for deployment is modular, designed to function in real-world environments. It has four components. The first is the user interface, the second is the inference engine, the third is the data management system, and the fourth is the AI chatbot module.

The primary interface with the system is the Graphical User Interface (GUI). In the web-based system, the user can input the CT scan of the brain or the clinical parameters. This is then sent to the model inference server, which then uses the EfficientNetV2-S and the FNN models to perform the prediction. The EfficientNetV2-S is used to perform the prediction with the CT scan, while the FNN is used with the clinical parameters.

There is also an AI chatbot module that interacts with the user in natural language to ask questions and educate the user about Alzheimer's disease. This module is known as the NeuroBot. The backend is designed to support an anonymized clinical dataset and to perform inference requests with the models.

The System deployment architecture of the proposed Alzheimer's detection framework is illustrated in Figure 3. The system integrates a web-based user interface with backend APIs, deep learning inference models (EfficientNetV2-S for brain CT slices and FNN for clinical data), and a chatbot module to support clinical decision-making and patient interaction.

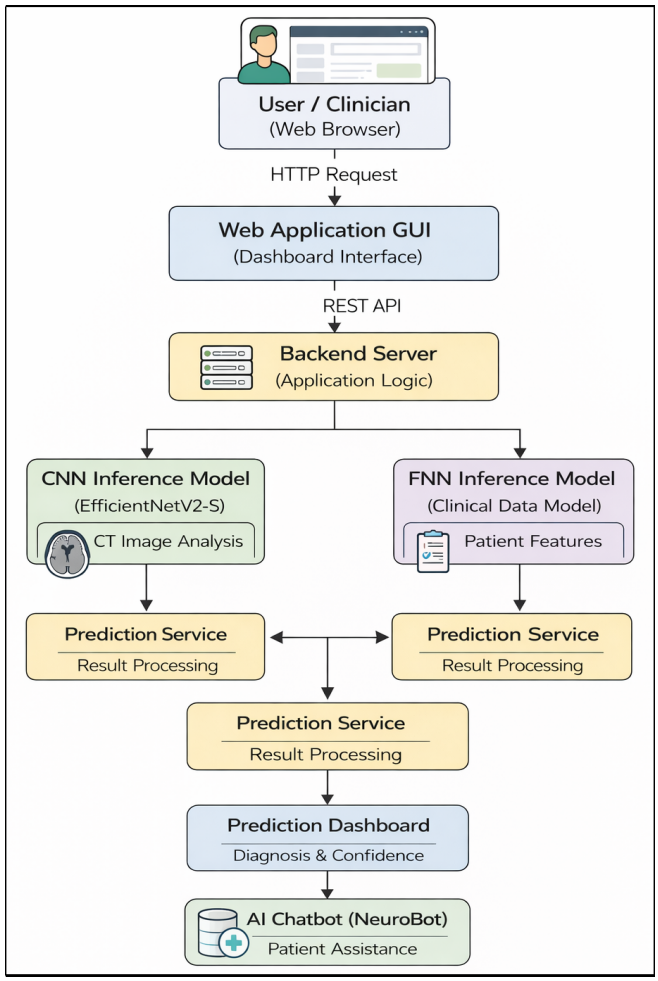


Figure 3. System Deployment Architecture Diagram.

3.3. Diagnostic Model Architectures

To create a comprehensive diagnostic tool, we employ two types of neural nets: Convolutional Neural Networks (CNNs) to process image data and Feedforward Neural Networks (FNNs) to process data based on clinical features.

3.3.1. Architectures for Image-Based Analysis (CNNs)

The fundamental operation in a CNN is the two-dimensional convolution. Convolution is a method of feature extraction, in which a kernel is applied to a given input image or feature map. This is arithmetically measured as:

$$O(i,j) = (I * K)(i,j) = \sum_m \sum_n I(i-m, j-n) \cdot K(m,n) \quad (1)$$

$O(i,j)$ characterises the output feature map of location (i,j), I is the input, and K is the convolutional kernel. Activation functionality is applied to add non-linearity. Our application is the ReLU, which is:

$$f(x) = \max(0, x) \quad (2)$$

For the image classification task, this study utilizes the EfficientNetV2-S architecture, a powerful and computationally efficient model. This model achieves its efficiency through the use of Fused-MBConv blocks in its early layers.

Fused-MBConv block simplifies the normal inverted residual block by fusing the first 3×3 depthwise convolution of the block with the second 1×1 projection convolution into a standard 3×3 convolution block. This minimizes the production of memory access overhead increases and enhances the training speed on advanced accelerators. The work of a Fused-MBConv block could be explained by the equation:

$$y = \text{BN}\left(\text{Conv}_{3\times3}\left(\text{BN}\left(\text{Conv}_{1\times1}(x)\right)\right)\right) + x \quad (3)$$

In this case, x is a signal sent into the block, and the equation shows the series of a 1×1 expansion convolution, a 3×3 standard convolution (a replacement of the individual depthwise and projection steps) and Batch Normalization (BN). The end result y is gotten by summing the original input x via a residual (skip) connection, which makes this a distinguishing feature that makes it likely to train very deep networks.

3.3.2. Architecture for Clinical Data Analysis (FNN)

A Feedforward Neural Network (FNN) was developed to analyze structured clinical data. This kind of network was trained using a hyperparameter optimization strategy that entailed the optimization of the predictive performance using the Optuna framework.

The FNN is made up of a series of dense (fully connected) layers. The general unit is its dense layer, which is a linear transform of its input vector x, which can be expressed as:

$$y = Wx + b \quad (4)$$

The input is represented by y, the learnable weight matrix is W, and the learnable bias is b.

The result of each of the dense layers is then run through the activation function of the Rectified Linear Unit (ReLU) to introduce non-linearity, as well as permit the model to learn more intricate patterns, and is defined as:

$$f(x) = \max(0, x) \quad (5)$$

Where γ and β are learnt scale and shift parameters, and the constant ϵ is a small number to maintain numerical stability. This is followed by the normalized output going through the (ReLU) activation function.

To avoid overfitting, a ReLU activation will be followed by a Dropout layer. Dropout is random and assigns a part of the input units to 0 at every update time throughout the training period, which assists in augmenting the strength of the network.

The optimized architecture that was discovered by hyperparameterOptimization is an input layer and two hidden layers:

1. The initial hidden layer is a dense layer that contains 57 units, an activation of ReLU, and a Dropout with a rate of 0.38.
2. The second hidden layer is a dense layer with 129 units, and the next layer consists of a ReLU activation and a Dropout layer at a rate of 0.49.

The concluding component is a single output neuron that produces a raw logit, z. This logit is then transformed into a probability using the sigmoid activation function. For improved numerical stability during training, this function is integrated directly into the loss function (BCEWithLogitsLoss). The sigmoid function is defined as:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \quad (6)$$

3.4. Interactive Framework and Training Algorithm

The project's workflow, executed within a Jupyter Notebook, provides two distinct diagnostic pathways: one for image-based analysis and another for clinical feature-based analysis. The classification pipeline for the image-based pathway is detailed in Algorithm 1, while the corresponding process for the clinical data FNN model is outlined in Algorithm 2.

Algorithm 1 AD Classification Pipeline (Image-Based CNN)

Input: Raw brain CT image I_{raw}.
Data Loading: The image dataset is loaded from the data/image directory and split into training and validation sets.
Preprocessing & Augmentation:
    $I_{\text{resized}} \Leftarrow \text{RandomResizedCrop}(I_{\text{raw}}, (224, 224))$
    $I_{\text{flipped}} \Leftarrow \text{RandomHorizontalFlip}(I_{\text{resized}})$
    $I_{\text{tensor}} \Leftarrow \text{ToTensor}(I_{\text{flipped}})$
    $I_{\text{norm}} \Leftarrow \text{Normalize}(I_{\text{tensor}})$
Model Loading: Load pre-trained model M (EfficientNetV2-S) and adapt its final layer for binary classification.
Training: The model is fine-tuned on the training set using an Adam optimizer and Binary Cross-Entropy with Logits loss (BCEWithLogitsLoss).
Prediction:
    $Z \Leftarrow M(I_{\text{norm}})$ // Get raw logit output from the CNN
Classification:
    $P \Leftarrow \sigma(Z)$ // Apply sigmoid function to get probability
    If $P > 0.5$, **then** $C \Leftarrow$ 'Demented'
    Else $C \Leftarrow$ 'Non-Demented'
Output: Return class label C.

Algorithm 2 Clinical Data Classification Pipeline (FNN)

Input: Raw clinical feature set D_{raw} from alzheimers_disease_data.csv.
Data Splitting: The dataset is split into training and testing sets (80/20 split).
Preprocessing:
    A ColumnTransformer pipeline is fitted on the training set:
        Numerical features are imputed (median) and standardized (StandardScaler).
        Categorical features are imputed (most frequent) and one-hot encoded (OneHotEncoder).
    $D_{\text{train\_processed}} \Leftarrow$ Apply fitted pipeline to training data.
    $D_{\text{test\_processed}} \Leftarrow$ Apply fitted pipeline to test data.
Class Imbalance Handling:
    $D_{\text{train\_resampled}} \Leftarrow$ Apply SMOTE (Synthetic Minority Over-sampling Technique) to $D_{\text{train\_processed}}$ to balance the classes.
Model Loading: Load the pre-trained Feedforward Neural Network (FNN).
Tensor Conversion:
    $T_{\text{input}} \Leftarrow \text{ToTensor}(D_{\text{test\_processed}})$
Prediction:
    $Z \Leftarrow M_{\text{fnn}}(T_{\text{input}})$ // Get raw logit output from the FNN
Classification:
    $P \Leftarrow \sigma(Z)$ // Apply sigmoid function to get probability
If $P > 0.5$, then $C \Leftarrow$ 'Demented'
Else $C \Leftarrow$ 'Non-Demented'
Output: Return class label C.

4. Results and Discussion

The efficacy of the hybrid framework that was suggested was critically evaluated using the aid of an integrated approach. This was accompanied by a quantitative measure of the classification performance of the deep learning models using a held-out test set of 4,500 brain CT slices, a large-scale measure of the associated clinical data to validate the validity of the features of the models, as well as a qualitative measure of the user interface and the interactivity aspects.

4.1. Performance Metrics

We used quantitative measures of the models using a sample set of standard classification measures that is decided by the elements of the confusion matrix: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

- **Accuracy:** The percentage of all the correct predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

- **Precision:** The proportion of positive predictions that were actually correct.

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

- **Recall (Sensitivity):** The proportion of actual positives that were identified correctly.

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

- **F1-Score:** The harmonic mean of Precision and Recall, providing a single score that balances both.

$$F1\text{-}Score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{10}$$

4.2. Model Interpretability and Performance Evaluation

Besides the model's accuracy, interpretability is another critical requirement for clinical decision-support system. In this study, the interpretability of the model was facilitated by the use of a technique called Gradient-weighted Class Activation Mapping (Grad-CAM) in the EfficientNetV2-S CNN model. This technique enables the production of visual maps that highlight areas of the images used in the model that are being used to make a prediction, as illustrated in Figure 4. These maps allow verification that the model focuses on anatomically relevant regions associated with Alzheimer's disease, such as cortical atrophy and ventricular enlargement.

Although Grad-CAM visualizations, statistical validation, and feature-importance analysis provide valuable insights into the model's decision-making process. Future work will investigate quantitative explainability metrics, including localization-based evaluation and clinician-assisted validation of explanation maps, to provide a more rigorous assessment of model interpretability and clinical trustworthiness.

Moreover, the interpretability of the clinical data pathway was facilitated by the use of statistical tests, such as independent sample t-tests and correlation tests, which showed that clinical variables, such as MMSE, Activities of Daily Living (ADL), and functional assessment, significantly vary across the different diagnostic groups.

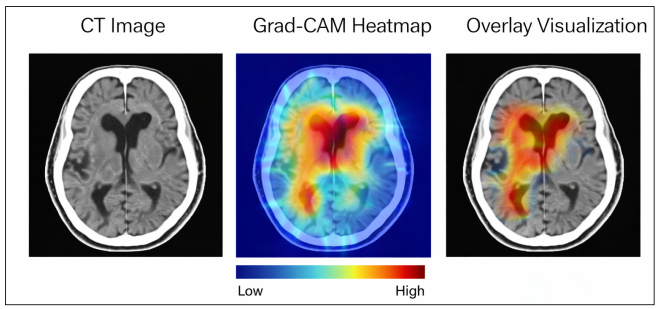


Figure 4. Grad-CAM visualization highlighting important regions in brain CT slices used by the EfficientNetV2-S model for Alzheimer's disease classification. The highlighted areas indicate regions contributing most strongly to the model's prediction.

The EfficientNetV2-S model achieved the best validation performance with an accuracy of 98.19% compared to other architectures, compared with other evaluated architectures, including EfficientNetV2-B0, ResNet50V2, and VGG16, as indicated in Table 1. Figure 5 shows the visualization of the performance of these models in comparison with each other, and it is evident that the EfficientNetV2-S model is more accurate.

Figure 6 displays the training and validation history of the EfficientNetV2-S model and depicts the change in the accuracy and loss after 10 epochs. This trend is convergently stable, and the trend has small overfitting, which is an attribute that indicates good fine-tuning and regularization.

Model	Accuracy (%)
EfficientNetV2-S	98.19%
EfficientNetV2-B0	95.8%
ResNet50V2	94.9%
VGG16	91.2%
DenseNet121	90.7%
MobileNetV2	89.3%

Table 1. Performance Comparison of Deep Learning Models

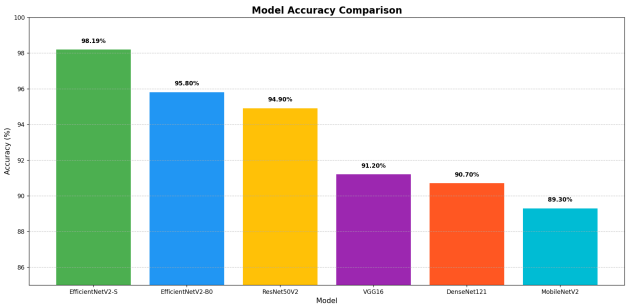


Figure 5. Visualization of Performance Comparison of All Models.

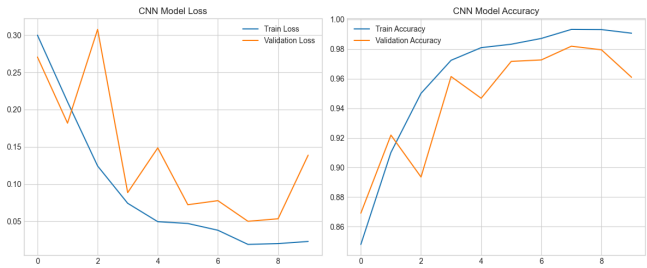


Figure 6. Training and Validation History for the EfficientNetV2-S Model, showing Loss and Accuracy over 10 epochs.

As a method of measuring the reliability of classifications, a confusion matrix of the EfficientNetV2-S model was constructed and presented in Figure 7. The matrix shows the accuracy of the model and recall of both the Demented and Non-Demented classes, that provide us with insight into the discriminatory ability of the model.

The FNN model, after undergoing hyperparameter tuning with Optuna, achieved a final test accuracy of 87.21% on the clinical dataset. This result validates the effectiveness of using structured clinical data for prediction and highlights the benefit of automated hyperparameter optimization.

In the dual-modality framework, there is the incorporation of medical images and clinical data, which makes the diagnosis even better. The CNN examines the images, digging deeper into the changes seen on the CT scan, such as cortical atrophy and enlarged ventricles, which indicate neurodegeneration. On the other hand, the FNN examines the clinical data, which includes cognitive tests such as MMSE, daily living activities, and other relevant demographic factors that may be associated with the condition. Statistical analysis reveals that MMSE, ADL, and other functional measures vary significantly across different diagnostic groups. The dual-modality framework, therefore, examines both visible changes on the brain images and the patient's individual risk factors, making the diagnosis even better and more reliable.

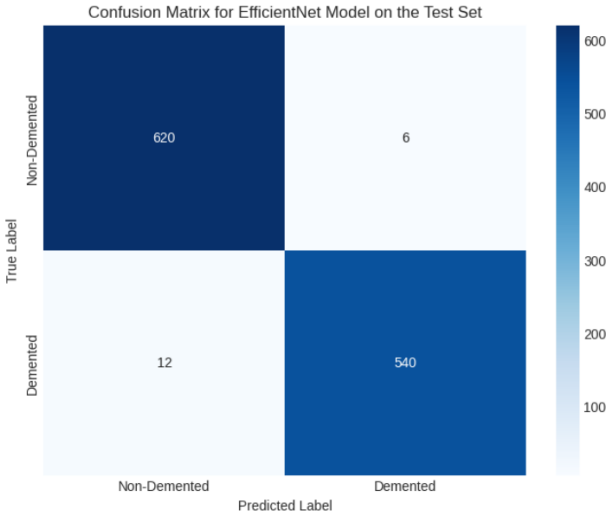


Figure 7. Confusion Matrix for EfficientNetV2-S on the Test Set.

To further evaluate the effectiveness of the proposed framework, its performance was compared with several recent state-of-the-art approaches for Alzheimer's detection reported in the literature. These studies employed various deep learning architectures and imaging modalities, including MRI-based convolutional neural networks and multimodal models. The comparison result presented in Table 2 demonstrates that the proposed framework achieves competitive performance while integrating CT imaging, structured clinical data, and an interactive clinical support system.

It should be noted that direct comparison of classification accuracies across studies should be interpreted with caution because the evaluated datasets, imaging modalities, sample sizes, and experimental protocols differ substantially. Most of the compared studies employed MRI-based datasets, particularly ADNI, whereas the proposed framework was developed using brain CT images and structured clinical data obtained from a hospital-based cohort. MRI generally provides higher soft-tissue contrast than CT imaging, which may influence diagnostic performance, Furthermore, variations in dataset composition, preprocessing procedures, and evaluation strategies can affect reported accuracies. Despite these differences, the proposed framework achieved competitive performance while simultaneously providing explainability through Grad-CAM and a deployment-oriented clinical support platform.

4.3. User Interface and System Interaction

The practical utility of the diagnostic models is realized through an intuitive and user-centric graphical user interface (GUI). The system's main dashboard, shown in Figure 8, provides a clean and accessible entry point for users. It presents two primary options for analysis, feature-based (clinical data) and image-based, alongside the integrated "NeuroBot" AI assistant.

For a feature-based analysis, the user navigates to a comprehensive data input form (Table 3), where they can enter demographic and clinical parameters. After submission, the system presents a detailed results page (Figure 9) that not only displays the diagnosis ('Demented' or 'Non-Demented') but also provides actionable suggestions and precautions tailored to the outcome. This aspect turns the system into a supportive system rather than a basic classifier. For an image-based diagnosis, the

Study	Data Modality	Model / Method	Dataset	Accuracy (%)
Pradhan et al., 2024	MRI	DenseNet + ResNet50	ADNI	95.3
Helaly et al., 2022	MRI	CNN-based Deep Learning	ADNI	93.6
Saleem et al., 2022	MRI	Transfer Learning CNN	ADNI	94.7
Ayus & Gupta, 2024	MRI	Hybrid Ensemble DL	ADNI	96.2
Alwakid et al., 2024	MRI	Image Processing + CNN	ADNI	94.5
Ávila-Jiménez et al., 2024	Clinical Records	Deep Learning Model	Clinical Dataset	85.0
Proposed Method	CT + Clinical Data	EfficientNetV2-S + FNN	Hospital + Clinical Dataset	98.19

Table 2. Comparison of the proposed method with recent state-of-the-art Alzheimer's disease detection approaches.

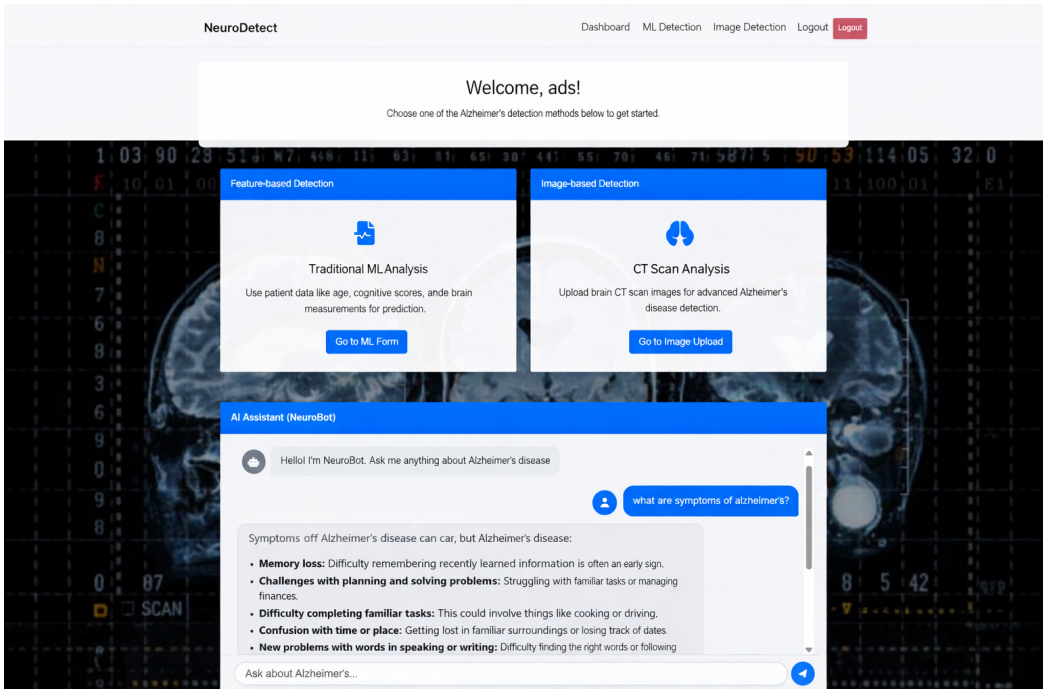


Figure 8. Main Dashboard Interface with Chatbot Panel.

user interacts with a simple drag-and-drop interface to upload a brain CT scan (Figure 10). The system processes the image and returns a clear results page displaying the diagnosis, a confidence score, and preventive measures or daily activities that may be beneficial (Figure 11). This consistent, informative clinical workflow is formulated to be easily ventured by clinicians, patients, and even caregivers, and this gap between a sophisticated AI model and a practical and usable tool is addressed.

4.4. Statistical Validation and Feature Analysis

To offer a sound clinical basis to our models, the structured clinical dataset underwent a complete statistical analysis. This validation guarantees that the characteristics learnt by the models are associated with accepted clinical markers of Alzheimer's disease.

This was done by the use of an independent samples t-test to identify the clinical characteristics that significantly differed between the Demented and Non-Demented groups. Figure 12 gives the general findings of the t-test in a manner that summarizes the comparative statistics of the two diagnostic groups. In the analysis, it was found that the significant indicators that constitute an overwhelming proportion of the significant indicators were the MMSE score, and that the p-value was 0.0000. These differences in distribution between the Age and MMSE age scores, as represented graphically in Figure 13 and Figure 14 of the boxplot is a very sharp and significant drop in MMSE scores in the demented group compared with Age.

The correlation heatmap in Figure 15 also shows correlations between major clinical characteristics and the final diagnosis, indicating that the MMSE score is most strongly negatively correlated with dementia diagnosis.

Most importantly, all these statistics are directly related to the feature-importance analysis of the FNN model presented in Figure 16. These characteristics, the FunctionalAssessment, ADL (Activities of Daily Living), MemoryComplaints and MMSE, were evaluated as the most effective predictors for the model. Such a high concordance rate assures that the FNN model evidently learns to focus on the clinically significant variables associated with Alzheimer's disease.

In addition to the visual explanations provided by Grad-CAM, quantitative evidence supporting model interpretability was obtained through statistical validation of the clinical variables. Independent sample t-tests demonstrated significant differences between demented and Non-demented groups, with MMSE exhibiting a highly significant association ($p < 0.001$). Correlation analysis further confirmed strong relationships between important clinical features and dementia

Feature	Value	Feature	Value
Age	68	SystolicBP	118
Gender	0	DiastolicBP	78
Ethnicity	0	CholesterolTotal	180
EducationLevel	3	CholesterolLDL	100
BMI	24.5	CholesterolHDL	65
Smoking	0	CholesterolTriglycerides	130
AlcoholConsumption	2	MMSE	29
PhysicalActivity	56	FunctionalAssessment	1.0
DietQuality	9	MemoryComplaints	0
SleepQuality	8	BehavioralProblems	0
FamilyHistoryAlzheimers	0	ADL	1.0
CardiovascularDisease	0	Confusion	0
Diabetes	0	Disorientation	0
Depression	0	Personality Changes	0
HeadInjury	0	Difficulty Completing Tasks	0
Hypertension	0	Forgetfulness	1
Doctor In Charge	221		

Table 3. Sample Input Values for Feature-Based Analysis.

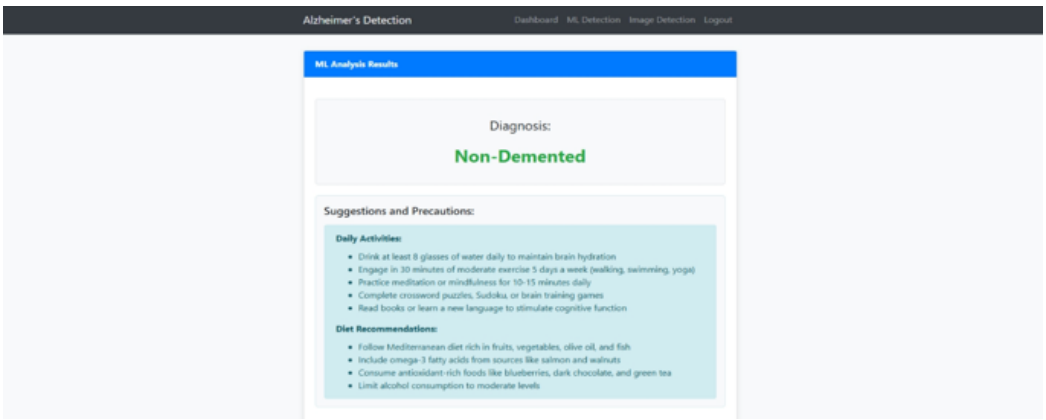


Figure 9. Analysis Results Page for Clinical Data Submission.

diagnosis. Furthermore, feature-importance analysis identified MMSE, Functional Assessment, ADL, and Memory Complaints as the most influential used by the FNN model. The agreement between statistical significance and model-derived feature importance provides quantitative evidence that the model focuses on clinically meaningful biomarkers associated with Alzheimer's disease.

4.5. User Engagement Framework

The final component of our evaluation focused on the AI-powered chatbot, NeuroBot, which serves as the primary tool for patient and caregiver support. The chatbot was assessed across five key criteria: Domain Relevance, Medical Accuracy, Politeness and Safety, Responsiveness, and its ability to refuse out-of-scope questions. As illustrated in the radar chart in Figure 17, NeuroBot scored perfectly or near-perfectly on all measures. This lends credibility to its accuracy, safety, and relevancy in delivering pertinent information within the field of Alzheimer Disease. Also, response time analysis revealed that the response to user queries was received in time (85% of queries were attended to within a range of 4.2 to 4.6 seconds), which guaranteed a consistent and interactive user experience.

5. Conclusion and Future Work

This study developed and evaluated a dual-modality framework for the early detection of Alzheimer's disease, where the deep learning pipeline is applied to the CT image processing and the hyperparameter-optimized Feedforward Neural Network (FNN) is applied to the clinical data processing.

The findings demonstrate that using a fine-tuned EfficientNetV2-S architecture for image analysis can achieve high diagnostic accuracy, reaching a peak validation performance of 98.19%. Complementing this, the FNN model, optimized through systematic hyperparameter tuning, achieved a final accuracy of 87.21% on structured clinical data. The statistical analysis of this clinical data further solidified our approach, confirming that the models are learning from features, such as the MMSE score, that are strongly correlated with established indicators of cognitive decline.

One of the major contributions that this research makes is the holistic approach. When putting these diagnostic models into a broader ecosystem, which includes a user-friendly interface and the AI-powered NeuroBot, this piece of work offers a blueprint of an all-inclusive support platform. This usability-based strategy is also essential in translating the gap between the complicated AI technology and the daily clinical practice,

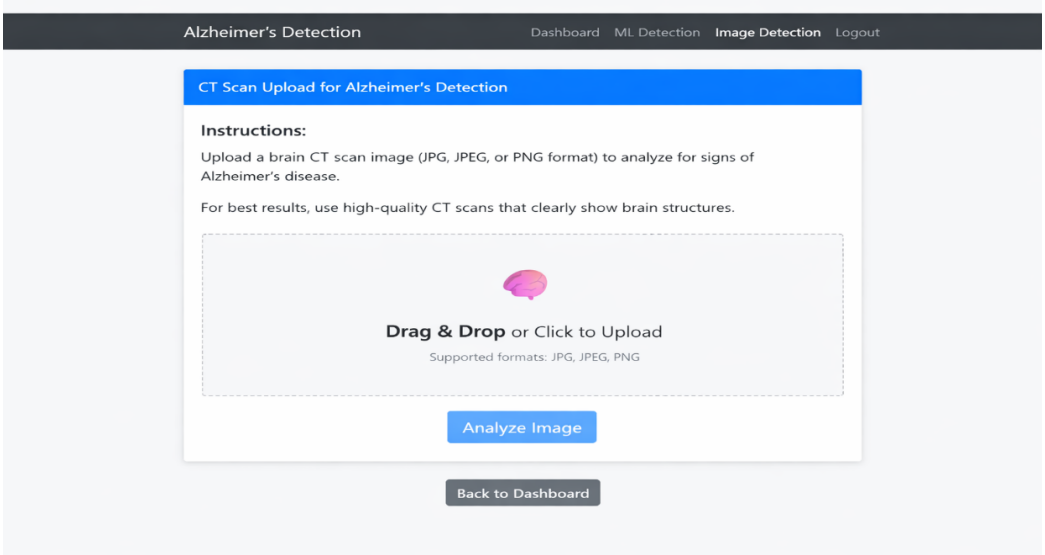


Figure 10. Drag-and-Drop Interface for Image-Based Analysis.

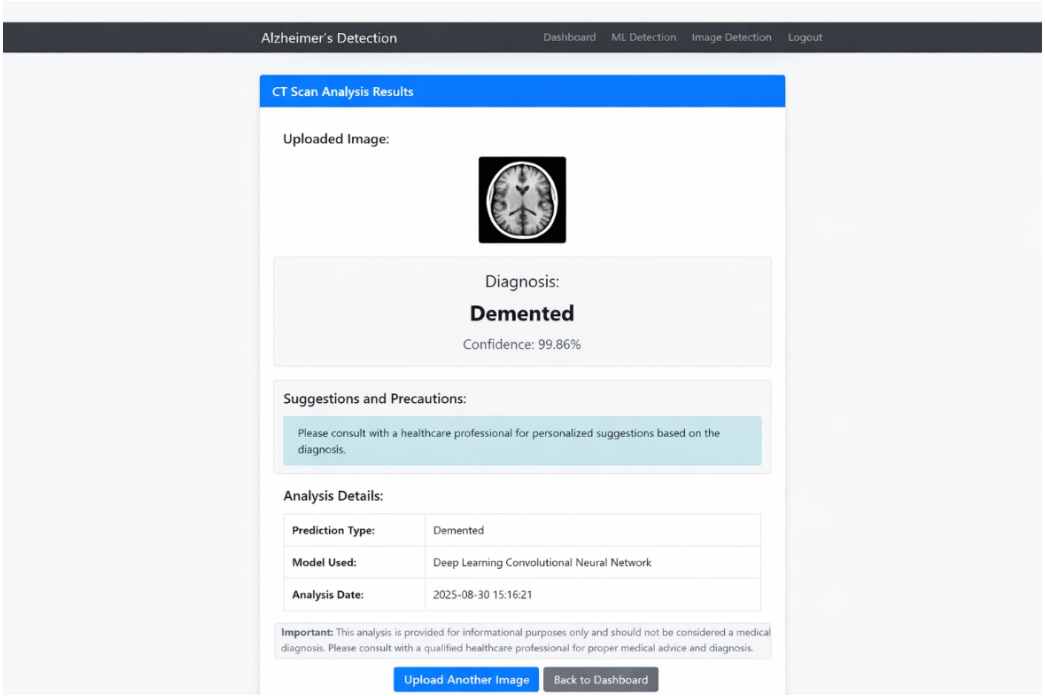


Figure 11. Analysis Results Page for Image-Based Diagnosis.

and allows patients, caregivers, and clinicians to place the right diagnostics and information right at their fingertips.

The proposed framework offers promising avenues for future development and presents several opportunities for improvement. The capabilities of the diagnostic functions will be expanded to provide a more comprehensive portrait of the patient, with a priority on the integration of multimodal data from MRI scans and genetic markers. We will also make the existing binary classification models extendable so that Alzheimer's can be classified in multiple steps, so that we are in a position to further distinguish between MCI and the onset of further stages of the disease.

The last and most important stage will be the push of the framework into clinical validation by means of large-scale trials. This will be essential for validating the system's performance across diverse populations and is a necessary step for its eventual integration into standard clinical workflows. Concurrently, we will explore the optimization of the models for deployment on edge devices and expand the chatbot's capabilities to include summarizing prediction results and offering multilingual support, thereby increasing the system's accessibility and global impact.

It should be noted that the above evaluation was carried out using controlled experimental data. Although the models were

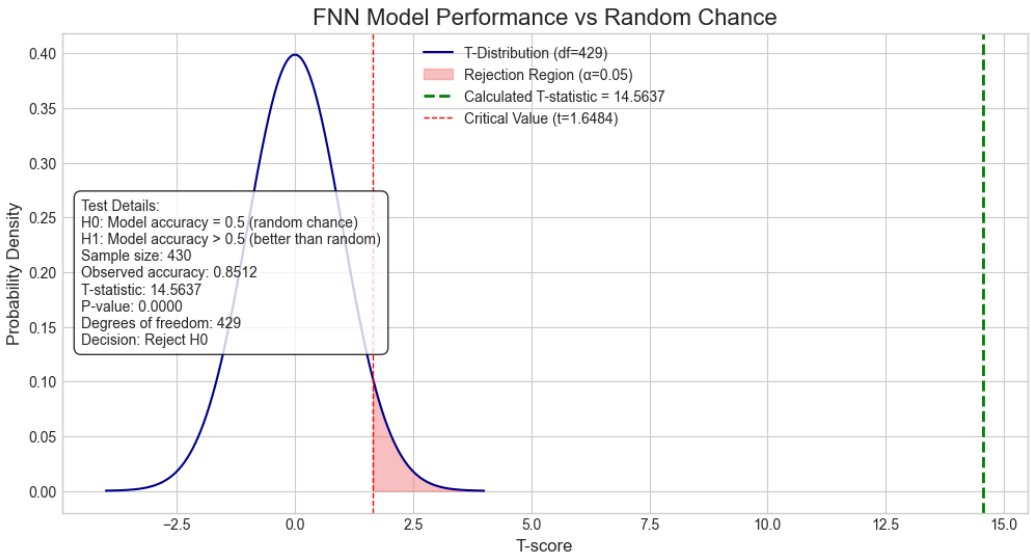


Figure 12. Visualization of T-Test Results Comparing Demented vs. Non-Demented Groups.

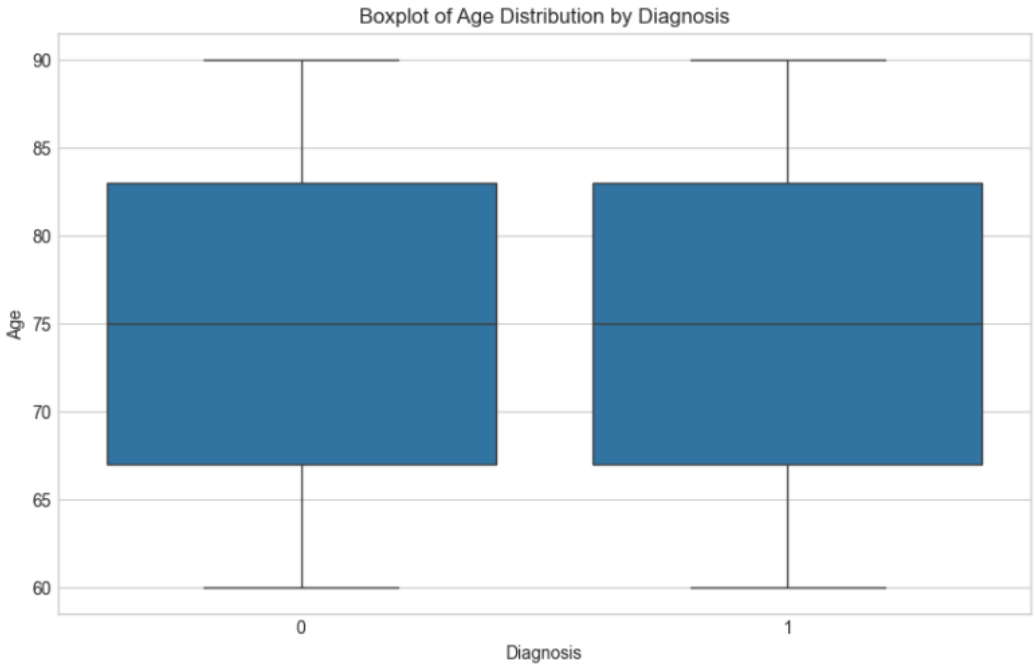


Figure 13. Boxplot of Age Distribution by Diagnosis.

validated using a held-out test set, the models were not validated using any independent clinical data sets. In the future, the focus will be on evaluating the framework using data sets from multiple institutions and clinical settings.

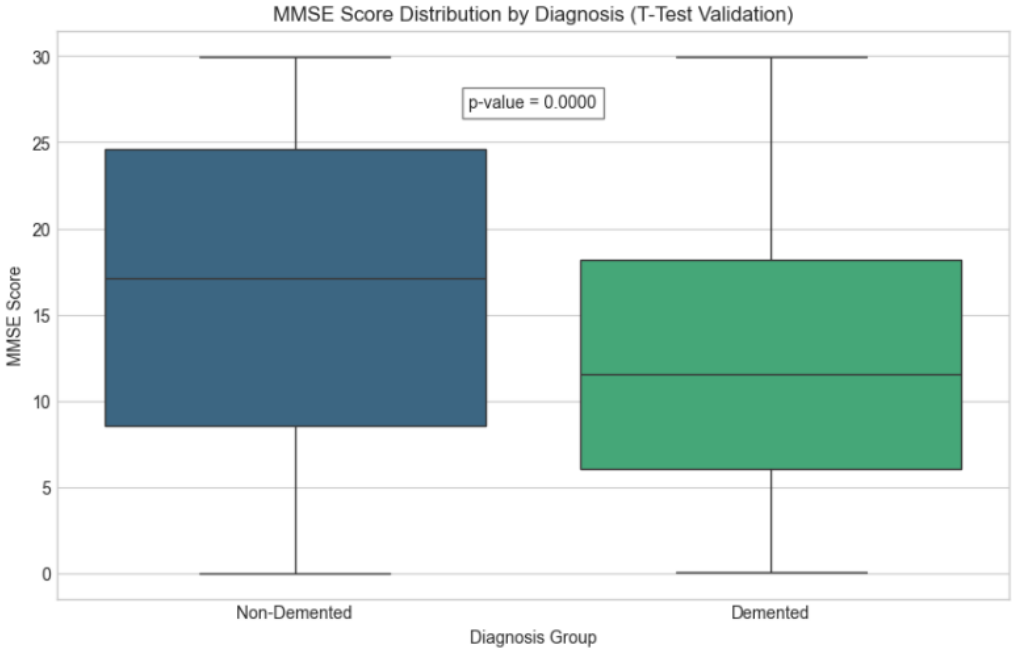


Figure 14. MMSE Score Distribution by Diagnosis (T-Test Validation).

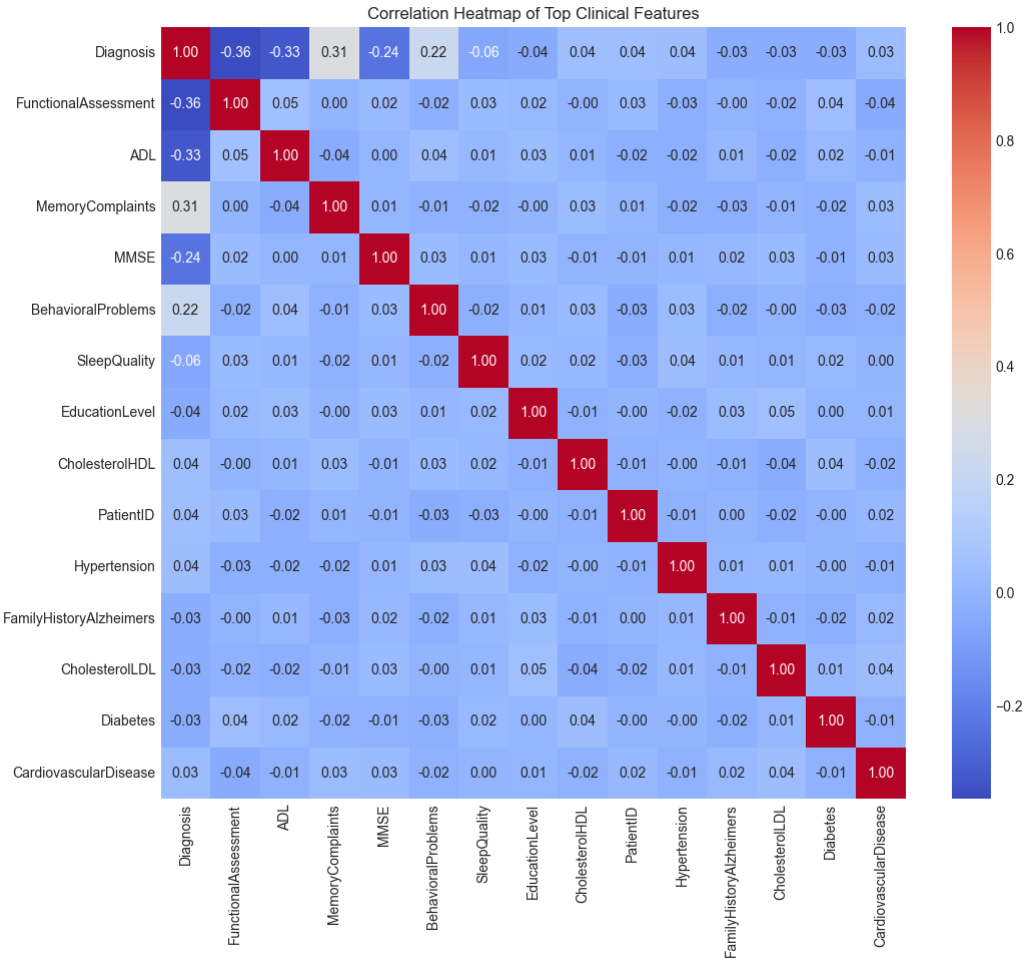


Figure 15. Correlation Heatmap of Top Clinical Features

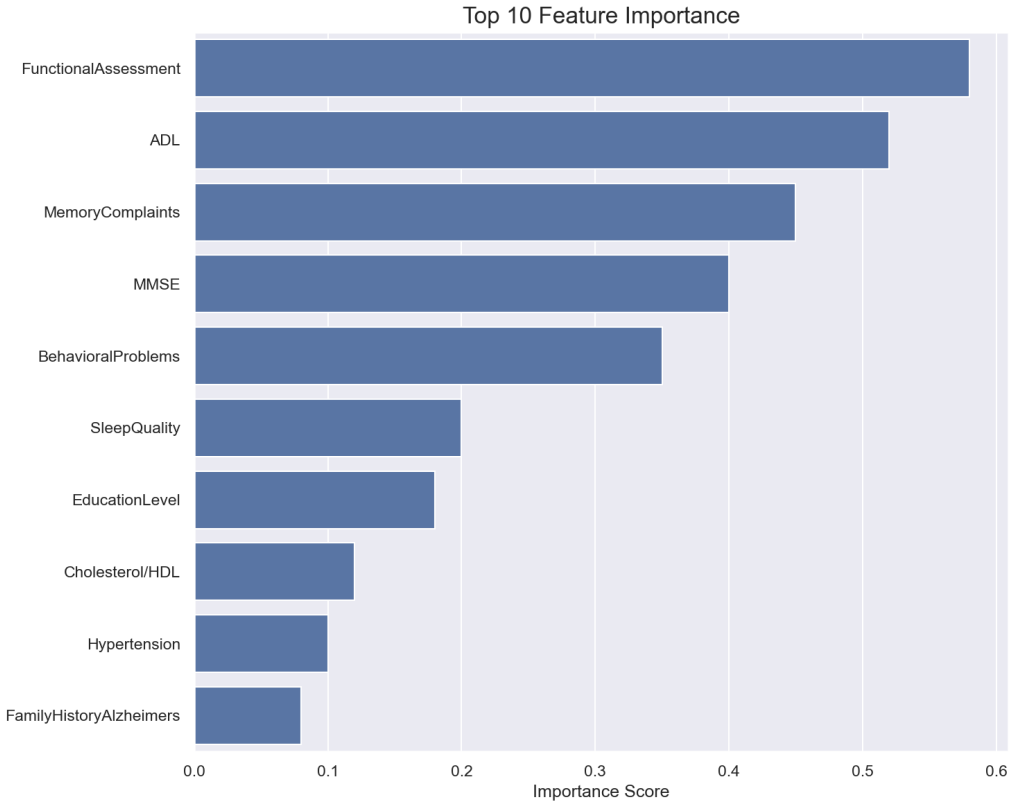


Figure 16. Top 10 Most Important Features.

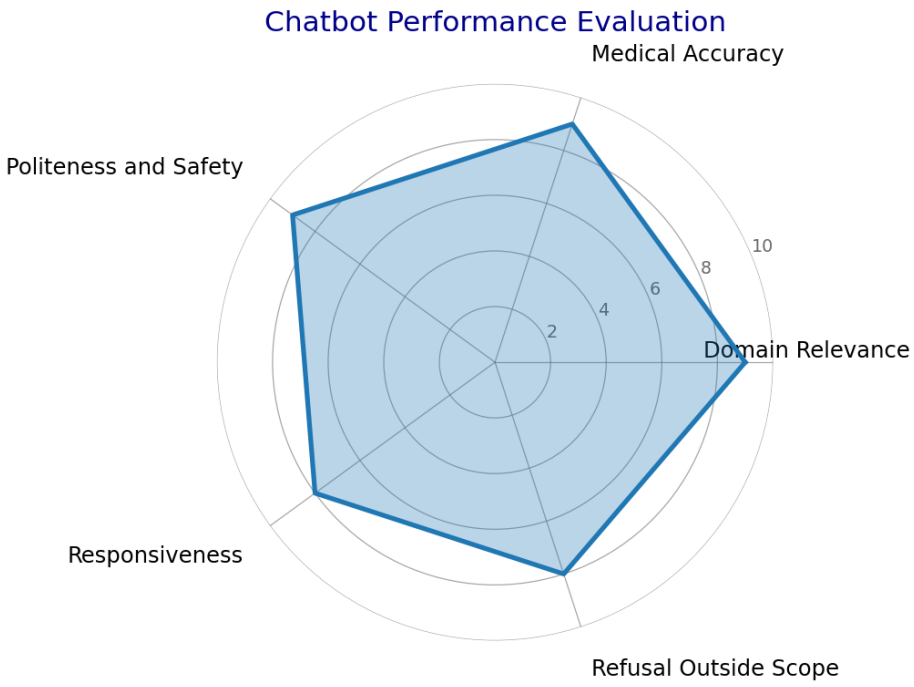


Figure 17. Radar Chart Evaluation of the Chatbot.

Ethical Statement

Ethical approval of this study was obtained from the Institutional Ethics Committee of Lovely Professional University, India (Ref: LPU/IEC-LPU/2025/1/2, dated 15 February 2025). Permission to access clinical and imaging data was granted by Shrimaan Superspeciality Hospital, Jalandhar, India. The study was retrospective and involved analysis of previously collected anonymised clinical and brain CT imaging data. No direct patient contact occurred, and no personally identifiable information was accessed. In accordance with the Institutional Ethics Committee clarification, the requirement for informed consent was waived/not applicable.

This study employed anonymized brain images from the brain CT slices of patients at Shrimann Superspeciality Hospital. No personally identifiable patient information was accessed during this study. The data were used strictly for academic research in accordance with the hospital's data privacy regulations. The proposed framework is intended for use as a clinical decision support tool. However, it is not intended for use as a substitute for clinical expertise. The NeuroBot chatbot provides general information guidance only and does not replace professional advice.

Funding

The scholar was sponsored by Tertiary Education Trust Fund (TETFUND) Nigeria for Higher education Studies only.

Declaration of competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability Statements

The data supporting the findings of this study are available from the corresponding author upon reasonable request. However, the data are not publicly available due to privacy or ethical restrictions.

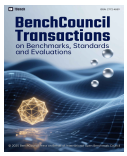
Credit authorship contribution statement

Shehu Mohammed: Conceptualization; Project Administration; Methodology; Data Curation; Software Development; Investigation; Writing – Original Draft Preparation.
Neha Malhotra: Supervision; Writing – Review & Editing; Formal Analysis.
Anmol Singh Rai: Resources; Validation; Visualization.

References

1. S. Alinsaif, "Dca-enhanced alzheimer's detection with shearlet and deep learning integration," *Computers in Biology and Medicine*, vol. 185, p. 109538, 2025, doi: https://doi.org/10.1016/j.compbiomed.2024.109538.
2. M. Srikanth and P. Bellapukonda, "The early detection of alzheimer's illness using machine learning and deep learning algorithms," *Journal of Pharmaceutical Negative Results*, vol. 13, no. 9, pp. 4852–4859, 2022, doi: https://doi.org/10.47750/pnr.2022.13.S09.603. [Online]. Available: https://www.pnrjournal.com/index.php/home/article/view/4470
3. E. M. Mohammed, A. M. Fakhrudeen, and O. Y. Alani, "Detection of alzheimer's disease using deep learning models: A systematic literature review," *Informatics in Medicine Unlocked*, vol. 50, p. 101551, 2024, doi: https://doi.org/10.1016/j.imu.2024.101551.
4. A. D. Arya *et al.*, "A systematic review on machine learning and deep learning techniques in the effective diagnosis of alzheimer's disease," *Brain Informatics*, vol. 10, no. 1, p. 17, 2023, doi: https://doi.org/10.1186/s40708-023-00195-7.
5. N. Pradhan, S. Sagar, and A. S. Singh, "Analysis of mri image data for alzheimer disease detection using deep learning techniques," *Multimedia Tools and Applications*, vol. 83, no. 6, pp. 17 729–17 752, 2024, doi: https://doi.org/10.1007/s11042-023-16256-2.
6. P. Kishore, C. U. Kumari, M. Kumar, and T. Pavani, "Detection and analysis of alzheimer's disease using various machine learning algorithms," *Materials today: proceedings*, vol. 45, pp. 1502–1508, 2021, doi: https://doi.org/10.1016/j.matpr.2020.07.645.
7. P. Balaji, M. A. Chaurasia, S. M. Bilfaqih, A. Muniasamy, and L. E. G. Alsid, "Hybridized deep learning approach for detecting alzheimer's disease," *Biomedicines*, vol. 11, no. 1, p. 149, 2023, doi: https://doi.org/10.3390/biomedicines11010149.
8. S. Mirabian, F. Mohammadian, Z. Ganji, H. Zare, and E. Hasanpour Khalesi, "The potential role of machine learning and deep learning in differential diagnosis of alzheimer's disease and ftd using imaging biomarkers: A review," *The Neuroradiology Journal*, vol. 38, no. 5, pp. 571–587, 2025, doi: https://doi.org/10.1177/19714009251313511.
9. J. Chua *et al.*, "Utilizing deep learning to predict alzheimer's disease and mild cognitive impairment with optical coherence tomography," *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, vol. 17, no. 1, p. e70041, 2025, doi: https://doi.org/10.1002/dad2.70041.
10. A. A. A. El-Latif, S. A. Chelloug, M. Alabdulhafith, and M. Hammad, "Accurate detection of alzheimer's disease using lightweight deep learning model on mri data," *Diagnostics*, vol. 13, no. 7, p. 1216, 2023, doi: https://doi.org/10.3390/diagnostics13071216.
11. K. Lokesh, N. P. Challa, A. S. Satwik, J. C. Kiran, N. K. Rao, and B. Naseeba, "Early alzheimer's disease detection using deep learning," *EAI Endorsed Transactions on Pervasive Health and Technology*, vol. 9, 2023, doi: https://doi.org/10.4108/eetpht.9.3966.
12. S. Balne and A. Elumalai, "Machine learning and deep learning algorithms used to diagnosis of alzheimer's," *Materials Today: Proceedings*, vol. 47, pp. 5151–5156, 2021, doi: https://doi.org/10.1016/j.matpr.2021.05.499.
13. S. Baskar, M. L. Prasad, N. Sharma, T. Katale, P. C. S. Reddy *et al.*, "An accurate prediction and diagnosis of alzheimer's disease using deep learning," in *2023 IEEE North Karnataka Subsection Flagship International Conference (NKCon)*. IEEE, 2023, pp. 1–7, doi: https://doi.org/10.1109/nkcon59507.2023.10396132.
14. A. Muydinov, "Advances in deep learning techniques for alzheimer's disease detection using mri images review," in *Proceedings of the 8th International Conference on Future Networks & Distributed Systems*, 2024, pp. 266–269, doi: https://doi.org/10.1145/3726122.3726160.

15. G. N. Alwakid, S. Tahir, M. Humayun, and W. Gouda, "Improving alzheimer's detection with deep learning and image processing techniques," *IEEE Access*, vol. 12, pp. 153 445–153 456, 2024, doi: https://doi.org/10.1109/ACCESS.2024.3481238.
16. G. S. Chhabra, A. Guru, B. J. Rajput, L. Dewangan, and S. K. Swarnkar, "Multimodal neuroimaging for early alzheimer's detection: a deep learning approach," in *2023 14th international conference on computing communication and networking technologies (ICCCNT)*. IEEE, 2023, pp. 1–5, doi: https://doi.org/10.1109/ICCCNT56998.2023.10307780.
17. S. Afzal *et al.*, "Alzheimer disease detection techniques and methods: A review." *International journal of interactive multimedia and artificial intelligence*, vol. 6, no. 7, pp. 26–38, 2021, doi: https://doi.org/10.9781/ijimai.2021.04.005.
18. S. Koga, A. Ikeda, and D. W. Dickson, "Deep learning-based model for diagnosing alzheimer's disease and tauopathies," *Neuropathology and Applied Neurobiology*, vol. 48, no. 1, p. e12759, 2022, doi: https://doi.org/10.1111/nan.12759.
19. I. Ayus and D. Gupta, "A novel hybrid ensemble based alzheimer's identification system using deep learning technique," *Biomedical Signal Processing and Control*, vol. 92, p. 106079, 2024, doi: https://doi.org/10.1016/j.bspc.2024.106079.
20. S. Suganyadevi, A. S. Rajasekaran, N. Satheesh, R. Suganthi, R. Naveenkumar *et al.*, "Alzheimer's disease diagnosis using deep learning approach," in *2023 Second International Conference on Electronics and Renewable Systems (ICEARS)*. IEEE, 2023, pp. 1205–1209, doi: https://doi.org/10.1109/ICEARS56392.2023.10085017.
21. J. L. Ávila-Jiménez, V. Cantón-Habas, M. del Pilar Carrera-González, M. Rich-Ruiz, and S. Ventura, "A deep learning model for alzheimer's disease diagnosis based on patient clinical records," *Computers in Biology and Medicine*, vol. 169, p. 107814, 2024, doi: https://doi.org/10.1016/j.compbiomed.2023.107814.
22. S. Savaş, "Detecting the stages of alzheimer's disease with pre-trained deep learning architectures," *Arabian Journal for Science and Engineering*, vol. 47, no. 2, pp. 2201–2218, 2022, doi: https://doi.org/10.1007/s13369-021-06131-3.
23. W. Al Shehri, "Alzheimer's disease diagnosis and classification using deep learning techniques," *PeerJ Computer Science*, vol. 8, p. e1177, 2022, doi: https://doi.org/10.7717/peerj-cs.1177.
24. T. J. Saleem *et al.*, "Deep learning-based diagnosis of alzheimer's disease," *Journal of Personalized Medicine*, vol. 12, no. 5, p. 815, 2022, doi: https://doi.org/10.3390/jpm12050815.
25. G. Hcini, I. Jdey, and H. Dhahri, "Investigating deep learning for early detection and decision-making in alzheimer's disease: A comprehensive review: G. hcini et al." *Neural Processing Letters*, vol. 56, no. 3, p. 153, 2024, doi: https://doi.org/10.1007/s11063-024-11600-5.
26. S. Jain and R. Srivastava, "Enhanced eeg-based alzheimer's disease detection using synchrosqueezing transform and deep transfer learning," *Neuroscience*, vol. 576, pp. 105–117, 2025, doi: https://doi.org/10.1016/j.neuroscience.2025.04.041.
27. H. A. Helaly, M. Badawy, and A. Y. Haikal, "Deep learning approach for early detection of alzheimer's disease," *Cognitive computation*, vol. 14, no. 5, pp. 1711–1727, 2022, doi: https://doi.org/10.1007/s12559-021-09946-2.
28. T. Vanaja, K. Shanmugavadivel, M. Subramanian, and C. Kanimozhiselvi, "Advancing alzheimer's detection: integrative approaches in mri analysis with traditional and deep learning models," *Neural Computing and Applications*, vol. 37, no. 14, pp. 8527–8546, 2025, doi: https://doi.org/10.1007/s00521-025-10993-1.
29. M. M. S. Fareed *et al.*, "Add-net: an effective deep learning model for early detection of alzheimer disease in mri scans," *IEEE Access*, vol. 10, pp. 96 930–96 951, 2022, doi: https://doi.org/10.1109/ACCESS.2022.3204395.
30. E. Kina, "Tleablcnn: Brain and alzheimer's disease detection using attention based explainable deep learning and smote using imbalanced brain mri," *IEEE Access*, 2025, doi: https://doi.org/10.1109/ACCESS.2025.3539550.
31. M. Taiyeb Khosroshahi *et al.*, "Explainable artificial intelligence in neuroimaging of alzheimer's disease," *Diagnostics*, vol. 15, no. 5, p. 612, 2025, doi: https://doi.org/10.3390/diagnostics15050612.



BenchCouncil Transactions on Benchmarks, Standards and Evaluations, 2026

DOI: https://doi.org/10.66834/ryt1gh36
Corrigendum

CORRIGENDUM

Corrigendum regarding missing Funding statements in previously published articles

Accepted on 20 June 2026

Abstract

Funding statements were missing in the published version of the following articles that appeared in previous issues of BenchCouncil Transactions on Benchmarks, Standards and Evaluations. The correct Funding statements, as provided and verified by the authors, are listed below.

1. "Predicting the number of call center incoming calls using deep learning" (BenchCouncil Transactions on Benchmarks, Standards and Evaluations Journal, 2025: 100213) https://doi.org/10.1016/j.tbench.2025.100213

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

2. "Comparative study of deep learning models for Parkinson's disease detection" (BenchCouncil Transactions on Benchmarks, Standards and Evaluations Journal, 2025: 100219) https://doi.org/10.1016/j.tbench.2025.100219

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

3. "An adaptive opposite slime mold feature selection algorithm for complex optimization problems" (BenchCouncil Transactions on Benchmarks, Standards and Evaluations Journal, 2025: 100250) https://doi.org/10.1016/j.tbench.2025.100250

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

4. "Evaluation of mechanical properties of natural fiber based polymer composite" (BenchCouncil Transactions on Benchmarks, Standards and Evaluations Journal, 2024: 100183) https://doi.org/10.1016/j.tbench.2024.100183

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

5. "MultiPoint: Enabling scalable pre-silicon performance evaluation for multi-task workloads" (BenchCouncil Transactions on Benchmarks, Standards and Evaluations Journal, 2024: 100189) https://doi.org/10.1016/j.tbench.2025.100189

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

6. "BinCodex: A comprehensive and multi-level dataset for evaluating binary code similarity detection techniques" (BenchCouncil Transactions on Benchmarks, Standards and Evaluations Journal, 2024: 100163) https://doi.org/10.1016/j.tbench.2024.100163

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

7. "Analyzing the impact of opportunistic maintenance optimization on manufacturing industries in Bangladesh: An empirical study" (BenchCouncil Transactions on Benchmarks, Standards and Evaluations Journal, 2024: 100172) https://doi.org/10.1016/j.tbench.2024.100172

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

8. "TensorTable: Extending PyTorch for mixed relational and linear algebra pipelines" (BenchCouncil Transactions on Benchmarks, Standards and Evaluations Journal, 2024: 100161) https://doi.org/10.1016/j.tbench.2024.100161

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

9. "Evaluatology: The science and engineering of evaluation" (BenchCouncil Transactions on Benchmarks, Standards and Evaluations Journal, 2024: 100162) https://doi.org/10.1016/j.tbench.2024.100162

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

10. "CloudAISim: A toolkit for modelling and simulation of modern applications in AI-driven cloud computing environments" (BenchCouncil Transactions on Benchmarks, Standards and Evaluations Journal, 2023: 100150) https://doi.org/10.1016/j.tbench.2024.100150

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

11. "Benchmarking, ethical alignment, and evaluation framework for conversational AI: Advancing responsible development of ChatGPT" (BenchCouncil Transactions on Benchmarks, Standards and Evaluations Journal, 2023: 100136) https://doi.org/10.1016/j.tbench.2023.100136

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

12. "Analyzing the potential benefits and use cases of ChatGPT as a tool for improving the efficiency and effectiveness of business operations" (BenchCouncil Transactions on Benchmarks, Standards and Evaluations Journal, 2023: 100140) https://doi.org/10.1016/j.tbench.2023.100140

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

13. "MetaverseBench: Instantiating and benchmarking metaverse challenges" (BenchCouncil Transactions on Benchmarks, Standards and Evaluations Journal, 2023: 100138) https://doi.org/10.1016/j.tbench.2023.100138

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

14. "Mind meets machine: Unravelling GPT-4's cognitive psychology" (BenchCouncil Transactions on Benchmarks, Standards and Evaluations Journal, 2023: 100139) https://doi.org/10.1016/j.tbench.2023.100139

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

15. "Unlocking the opportunities through ChatGPT Tool towards ameliorating the education system" (BenchCouncil Transactions on Benchmarks, Standards and Evaluations Journal, 2023: 100115) https://doi.org/10.1016/j.tbench.2023.100115

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

16. "Benchmarking HTAP databases for performance isolation and real-time analytics" (BenchCouncil Transactions on Benchmarks, Standards and Evaluations Journal, 2023: 100122) https://doi.org/10.1016/j.tbench.2023.100122

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

17. "CoviDetector: A transfer learning-based semi supervised approach to detect Covid-19 using CXR images" (BenchCouncil Transactions on Benchmarks, Standards and Evaluations Journal, 2023: 100119) https://doi.org/10.1016/j.tbench.2023.100119

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

18. "DPUBench: An application-driven scalable benchmark suite for comprehensive DPU evaluation" (BenchCouncil Transactions on Benchmarks, Standards and Evaluations Journal, 2023: 100120) https://doi.org/10.1016/j.tbench.2023.100120

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

19. "ERMDS: A obfuscation dataset for evaluating robustness of learning-based malware detection system" (BenchCouncil Transactions on Benchmarks, Standards and Evaluations Journal, 2023: 100106) https://doi.org/10.1016/j.tbench.2023.100106

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

20. "SNNBench: End-to-end AI-oriented spiking neural network benchmarking" (BenchCouncil Transactions on Benchmarks, Standards and Evaluations Journal, 2023: 100108) https://doi.org/10.1016/j.tbench.2023.100108

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

21. "Enabling hyperscale web services" (BenchCouncil Transactions on Benchmarks, Standards and Evaluations Journal, 2023: 100192) https://doi.org/10.1016/j.tbench.2023.100092

Funding: This work was supported by (1) the Center for Applications Driving Architectures (ADA), one of six centers of JUMP, a Semiconductor Research Corporation program cosponsored by DARPA; (2) NSF Grant IIS1539011; (3) gifts from Intel and Google; and (4) a Facebook Fellowship.

22. "ChatGPT for healthcare services: An emerging stage for an innovative perspective" (BenchCouncil Transactions on Benchmarks, Standards and Evaluations Journal, 2023: 100105) https://doi.org/10.1016/j.tbench.2023.100105

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

23. "HPC AI500 V3.0: A scalable HPC AI benchmarking framework" (BenchCouncil Transactions on Benchmarks, Standards and Evaluations Journal, 2022: 100083) https://doi.org/10.1016/j.tbench.2022.100083

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

24. "Performance characterization and optimization of pruning patterns for sparse DNN inference" (BenchCouncil Transactions on Benchmarks, Standards and Evaluations Journal, 2022: 100090) https://doi.org/10.1016/j.tbench.2023.100090

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

25. "Enabling Reduced Simpoint Size Through LiveCache and Detail Warmup" (BenchCouncil Transactions on Benchmarks, Standards and Evaluations Journal, 2022: 100082) https://doi.org/10.1016/j.tbench.2022.100082

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

26. An era of ChatGPT as a significant futuristic support tool: A study on features, abilities, and challenges" (BenchCouncil Transactions on Benchmarks, Standards and Evaluations Journal, 2022: 100089) https://doi.org/10.1016/j.tbench.2023.100089

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

27. "An extensive study on Internet of Behavior (IoB) enabled Healthcare-Systems: Features, facilitators, and challenges" (BenchCouncil Transactions on Benchmarks, Standards and Evaluations Journal, 2022: 100085) https://doi.org/10.1016/j.tbench.2023.100085

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

28. "High fusion computers: The IoTs, edges, data centers, and humans-in-the-loop as a computer" (BenchCouncil Transactions on Benchmarks, Standards and Evaluations Journal, 2022: 100075) https://doi.org/10.1016/j.tbench.2022.100075

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

29. "A review of Blockchain Technology applications for financial services" (BenchCouncil Transactions on Benchmarks, Standards and Evaluations Journal, 2022: 100073) https://doi.org/10.1016/j.tbench.2022.100073

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

30. "SAIBench: Benchmarking AI for Science" (BenchCouncil Transactions on Benchmarks, Standards and Evaluations Journal, 2022: 100063) https://doi.org/10.1016/j.tbench.2022.100063

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

31. "Performance and energy consumption tradeoff in server consolidation" (BenchCouncil Transactions on Benchmarks, Standards and Evaluations Journal, 2022: 100060) https://doi.org/10.1016/j.tbench.2022.100060

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

32. "Challenges and recent advances in the design of real-time wireless Cyber-Physical Systems" (BenchCouncil Transactions on Benchmarks, Standards and Evaluations Journal, 2022: 100036) https://doi.org/10.1016/j.tbench.2022.100036

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

33. "Benchmarking feature selection methods with different prediction models on large-scale healthcare event data" (BenchCouncil Transactions on Benchmarks, Standards and Evaluations Journal, 2021: 100004) https://doi.org/10.1016/j.tbench.2021.100004

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

34. "MVDI25K: A large-scale dataset of microscopic vaginal discharge images" (BenchCouncil Transactions on Benchmarks, Standards and Evaluations Journal, 2021: 100008) https://doi.org/10.1016/j.tbench.2021.100008

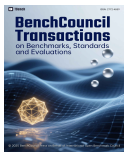
Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

35. "Fallout: Distributed systems testing as a service" (BenchCouncil Transactions on Benchmarks, Standards and Evaluations Journal, 2021: 100010) https://doi.org/10.1016/j.tbench.2021.100010

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

36. "Revisiting the effects of the Spectre and Meltdown patches using the top-down microarchitectural method and purchasing power parity theory" (BenchCouncil Transactions on Benchmarks, Standards and Evaluations Journal, 2021: 100011) https://doi.org/10.1016/j.tbench.2021.100011

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.



BenchCouncil Transactions on Benchmarks, Standards and Evaluations, 2026

DOI: https://doi.org/10.66834/t02ph695
Corrigendum

CORRIGENDUM

Corrigendum regarding incorrect Declaration of Conflict-of-Interest Statements in Previously Published Articles

Accepted on 11 June 2026

Abstract

Declaration of Competing Interest statements were incorrectly included in the published version of the following articles that appeared in previous issues of BenchCouncil Transactions on Benchmarks, Standards and Evaluations. The appropriate Conflict-of-Interest Statements, provided by the authors, are included below.

1."Evaluatology-driven artificial intelligence"(BenchCouncil Transactions on Benchmarks, Standards and Evaluations Journal, 2025: 100245) https://doi.org/10.1016/j.tbench.2025.100245

Declaration of competing interest: Guoxin Kang is an Associate Editor, Wanling Gao is an Assistant Editor-in-Chief, and Jianfeng Zhan is the Editor-in-Chief of BenchCouncil Transactions on Benchmarks, Standards and Evaluations. They were not involved in the editorial review process or the decision to publish this article.

2."MultiPoint: Enabling scalable pre-silicon performance evaluation for multi-task workloads"(BenchCouncil Transactions on Benchmarks, Standards and Evaluations Journal, 2024: 100189) https://doi.org/10.1016/j.tbench.2025.100189

The authors Yuxuan Wu and Wenxiang Wang declare the following personal relationship that may be considered a potential competing interest: they are currently employed by Loongson Technology, Beijing, China.