BenchCouncil Transactions
on Benchmarks, Standards and Evaluations

TBench
Volume 5, Issue 3
2025

Original Articles

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.

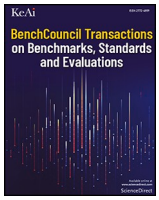
Contents

BenchCouncil Transactions on Benchmarks, Standards and Evaluations 5 (2025) 100228

Contents lists available at ScienceDirect

BenchCouncil Transactions on Benchmarks, Standards and Evaluations

journal homepage: www.keaipublishing.com/en/journals/benchcouncil-transactions-onbenchmarks-standards-and-evaluations/



Research Article

Exposing financial shenanigans: The role of Indian accounting standards (Ind AS) in enhancing corporate accountability and governance

Sunil Kumar

Faculty of Management & Commerce, ICFAI University Tripura, Tripura, India

ARTICLE INFO

Keywords:
Indian accounting standards
Financial shenanigans
Corporate governance
Financial reporting
Investor protection
Corporate accountability

ABSTRACT

The Indian Accounting Standards (Ind AS) play a pivotal role in reducing financial impropriety. These standards significantly enhance the accountability, accuracy, and transparency of financial reporting, thereby serving an essential function in deterring financial malfeasance. Such malfeasance includes deceptive accounting practices, misleading reporting, and the distortion of earnings, all of which undermine investor confidence, disrupt market integrity, and adversely affect the economy. The Ind AS, aligned with the International Financial Reporting Standards (IFRS), provide a comprehensive and robust framework that substantially improves the quality of financial reporting. The article outlines the significant benefits of Ind AS for financial reporting, such as increased transparency and accuracy. It presents case studies illustrating how the application of the standard has effectively addressed and mitigated financial discrepancies. Furthermore, the article examines the challenges organisations face in adopting Ind AS, including the complexities of transitioning from previous accounting standards and the need for extensive system reforms and personnel training. By elucidating these challenges, the article offers a thorough analysis of the effectiveness of Ind AS in addressing financial malpractice. It emphasises its role in fostering a more transparent and responsible financial reporting environment.

1. Introduction

Financial shenanigans—practices designed to manipulate financial statements and present a distorted view of a company's performance—pose significant risks to stakeholders and undermine the overall integrity of financial markets [1,2]. In response to the rising incidence of corporate fraud and the demand for greater transparency, the Indian Accounting Standards (Ind AS) were introduced to ensure consistency, accuracy, and fairness in financial reporting. Based on the International Financial Reporting Standards (IFRS), Ind AS aims to establish uniformity across sectors and provide investors with a clear and truthful representation of companies' financial health [3].

Ind AS plays a pivotal role in curbing financial shenanigans by enforcing stringent disclosure norms, promoting fair value measurement, and requiring detailed transaction reporting. Through standards like Ind AS 1 (Presentation of Financial Statements), Ind AS 109 (Financial Instruments), and Ind AS 115 (Revenue from Contracts with Customers), Indian firms are obligated to present their financial statements in a manner that reduces the scope for manipulation, fraudulent revenue recognition, and asset misrepresentation [4,5]

By mandating transparent financial disclosures and reinforcing corporate governance, Ind AS has become a vital tool for auditors, regulators, and investors to identify and prevent financial irregularities. This system improves financial discipline within companies and fosters trust in the Indian capital markets [6].

1.1. Literature review

Financial shenanigans, the gimmicks used to misrepresent financial statements through questionable accounting practices, continue to pose a significant global challenge to investors and regulators [7]. However, developing more stringent financial reporting standards does not seem to have effectively curbed these unethical practices, which persist worldwide [8]. The search for potential factors that may lead companies to engage in such unethical behaviour has been a primary motivation behind recent research in this domain.

This literature review examines the existing body of research on accounting fraud, with a focus on the role of Indian Accounting Standards in addressing this issue [9]. The findings highlight the importance of responsible corporate governance, sound accounting practices, and

Peer review under the responsibility of The International Open Benchmark Council.

E-mail address: sunilkumar@iutripura.edu.in.

https://doi.org/10.1016/j.tbench.2025.100228

Available online 11 July 2025

the psychological characteristics of managers and employees as critical factors influencing the prevalence of unethical behaviour. These psychological characteristics may include greed, pressure to meet financial targets, and the fear of job loss [8].

Companies have long employed various deceptive accounting practices, collectively known as "financial shenanigans," to misrepresent their trustworthy financial standing and mislead stakeholders [7]. Despite the development of more demanding financial reporting standards, the problem of accounting fraud persists worldwide, suggesting that the existing measures may not effectively address this issue [8]. To enhance transparency and accountability in the financial reporting practices of Indian companies, the government has introduced the Indian Accounting Standards, which closely align with the International Financial Reporting Standards adopted globally [7,8,10]. This literature review aims to synthesise the current scholarly understanding of the role and effectiveness of these standards in mitigating the prevalence of financial shenanigans in the Indian corporate sector. The body of academic research has consistently demonstrated the far-reaching implications of accounting fraud, which can undermine the confidence of individual investors and creditors and jeopardise the overall stability and growth of the economy [7,8]. Scholars have emphasised the need for a multifaceted approach to address this challenge, one that combines strengthened regulatory oversight, the implementation of robust corporate governance practices, and the cultivation of enhanced ethical awareness and integrity among financial professionals

The Indian Accounting Standards (Ind-AS) play a pivotal role in preventing financial fraud by enhancing the quality and reliability of financial reporting. This fosters transparency and accountability among businesses, reassuring stakeholders and the public. The transition from Generally Accepted Accounting Principles (GAAP) to International Financial Reporting Standards (IFRS) aimed to align Indian practices with global standards, which is essential for attracting foreign investment and ensuring comparability in financial statements [11]. By mandating rigorous disclosure requirements and measurement criteria, Ind-AS compels companies to adopt robust internal controls and governance structures, which are vital in mitigating fraud risks [11]. The involvement of auditors and forensic accountants in the financial reporting process is critical, as they utilise these standards to detect and prevent fraudulent activities through comprehensive audits and investigations [12,13–14]. Overall, the implementation of Ind-AS not only enhances financial integrity but also empowers stakeholders to identify and address potential fraud effectively.

The literature strongly suggests that the social cost of accounting fraud should be minimised. Governments and companies must urgently develop policies combining responsible corporate governance with environmental sustainability. This is not just a financial issue but a societal one, requiring immediate attention and action.

This discussion is crucially concerned with the role of Indian Accounting Standards in curbing financial shenanigans. These standards ensure transparent and accurate financial reporting, reducing the scope for manipulation and misrepresentation [1].

1.2. Research objectives

1. Assess how Ind AS, with its potential to enhance the accuracy, transparency, and reliability of financial reports, can significantly reduce financial misreporting and foster a more transparent and reliable financial landscape.
2. Explore how the adoption of Ind AS can revolutionise corporate governance by enforcing stricter disclosure norms and accountability, and enhancing transparency and ethical behaviour.
3. Conduct a thorough analysis of the practical challenges businesses may face during the implementation of Ind AS, including system overhauls and employee training, to provide a realistic view of the process.
4. Investigate how the adoption of Ind AS enhances investor protection by promoting transparency and ethical corporate behaviour.
5. Evaluate case studies of companies adopting Ind AS to reduce financial shenanigans and fraudulent reporting.

1.3. Research hypotheses

1. The adoption of Ind AS significantly reduces financial shenanigans by enhancing financial reporting accuracy, transparency, and accountability (H1).
2. There is no significant relationship between the adoption of Ind AS and the reduction of financial malfeasance in financial reporting (H0).
3. Ind AS improves corporate governance and investor protection by enforcing stricter disclosure norms and ethical behaviour (H2).
4. Challenges in adopting Ind AS, such as system changes and employee training, negatively impact its implementation effectiveness (H3).
5. Companies that successfully implement Ind AS experience fewer financial anomalies than those that do not (H4).

1.4. Research methodology

This study employs a strong mixed-methods research design that integrates quantitative analysis with qualitative case-based examination, centring on the impacts of Indian Accounting Standards (Ind AS) on financial transparency, misreporting risks, the pandemic, and corporate governance practices [15]. The quantitative component employs a longitudinal framework based on panel data, utilising descriptive statistics, independent sample *t*-tests, and multinomial logistic regression, along with model fit diagnostics to ensure both statistical robustness and empirical validity [16].

This study explores the financial performance of companies before and after implementing Ind AS, assesses cases of possible financial misreporting, analyses the evolution of corporate governance structures, and considers the external impacts of the COVID-19 pandemic using dummy variables [17].

1.5. Data analysis techniques

A wide range of both quantitative and qualitative methods was employed to assess the impact of implementing Ind AS on financial reporting practices in India [18]. The quantitative analysis includes descriptive statistics, inferential tests using *t*-tests, and predictive modelling with multinomial logistic regression. To ensure the validity, reliability, and explanatory strength of the models, diagnostics such as goodness-of-fit measures were applied [19,20].

To assess financial transparency, risk exposure, and resilience, a detailed examination of key financial ratios was conducted across pre- and post-Ind AS, Suspected Fraud, Covid-19 pandemic, and the Emergence of corporate governance as dependent dummy variables in implementation periods to identify shifts in performance metrics and reporting accuracy [21].

A comparative longitudinal analysis was also conducted to examine the changes in financial disclosure practices and governance frameworks resulting from the adoption of specific Indian Accounting Standards (Ind AS) provisions. The qualitative part utilised a case study approach with content analysis to investigate context-relevant factors affecting standards adoption. The case studies focused on the implementation of selected Ind AS norms—specifically, Ind AS 1, 24, 36, 37, 109, 110, and 115 [22].

1.6. Analytical tools utilised

Data analysis and econometric modelling were performed using Microsoft Excel, IBM SPSS Statistics, Gretl, and R Studio. These tools enabled data cleaning, statistical calculations, and the use of advanced

econometric methods. Their integrated application guaranteed analytical rigour, reproducibility, and the integrity of data-driven insights throughout all stages of the research.

1.7. Scope of the study

This study encompasses the periods preceding and following the implementation of Ind AS in the Indian corporate sector. It aims to assess the transition to Ind AS by examining its impact on financial transparency, reduction of financial fraud, strengthening corporate governance frameworks, and boosting investor confidence [23]. Utilising a firm-level empirical approach complemented by regulatory insights, this research makes a significant contribution to the discussion on accounting reform and financial governance in emerging markets.

1.8. Expected outcomes

1. Demonstrate that Ind AS improves financial reporting accuracy and reduces financial fraud.
2. Demonstrate that companies with robust governance structures benefit more from adopting Ind AS.
3. Highlight challenges in Ind AS implementation and offer solutions.
4. Conclude that Ind AS enhances investor confidence by promoting transparency and ethical behaviour.
5. Provide case study evidence showing that Ind AS reduces financial anomalies and promotes corporate ethics.

2. Content analysis: Indian accounting standards in preventing financial frauds

Indian Accounting Standards (Ind AS) play a significant role in preventing financial fraud. They ensure transparency, consistency, and accountability in financial reporting [1]. These standards, which have converged with the International Financial Reporting Standards (IFRS) to form a robust framework, help organisations present their financial statements honestly and fairly [3]. This convergence with IFRS means that Ind AS is not just a local standard but aligns with global best practices in financial reporting. Ind AS is a powerful tool in the fight against fraud [24].

2.1. Enhanced transparency and disclosure requirements

Ind AS mandates extensive disclosures, making it difficult for companies to hide or misrepresent financial information. By requiring detailed notes on various financial aspects, such as revenue recognition, related-party transactions, and financial instruments, the standards reduce opportunities for manipulation [25].

2.2. Fair valuation and measurement

Ind AS emphasises fair value measurement for assets and liabilities instead of historical cost accounting. This approach limits the ability to inflate or undervalue assets, provides a realistic financial picture, and prevents asset overstatement or understatement fraud.

2.3. Revenue recognition (Ind AS 115)

The recognition standard under Ind AS outlines stringent principles for recognising revenue, preventing companies from prematurely booking revenues to inflate earnings. This deters fraud involving the manipulation of sales figures or earnings reports [3].

2.4. Accounting for financial instruments (Ind AS 109)

Ind AS 109 ensures accurate classification and measurement of financial assets and liabilities, including provisions for expected credit losses. This helps prevent the concealment of bad debts and ensures financial institutions accurately report their credit risk exposures.

2.5. Consolidation of financial statements (Ind AS 110)

Ind AS 110 mandates the consolidation of financial statements for subsidiaries and other controlled entities. This eliminates the possibility of hiding liabilities or manipulating the financial performance of group companies.

2.6. Stringent corporate governance

The standards enhance corporate governance by encouraging the establishment of internal controls and risk management systems aligned with accounting practices. This reduces opportunities for fraud and helps in early detection [25].

2.7. Auditor's role

With the implementation of Ind AS, auditors are required to pay closer attention to compliance with these standards. This ensures that financial statements are scrutinised more rigorously, making it harder for companies to commit fraud without detection [25].

2.8. Consistency and comparability

By standardising accounting practices, Ind AS ensures consistency across periods and company comparability. This uniformity reduces the scope for manipulation through inconsistent accounting treatments [25].

3. Indian accounting standards (Ind AS) help in preventing financial fraud

Indian Accounting Standards (Ind AS) enhance transparency, consistency, and comparability in financial reporting, thereby minimising opportunities for manipulation and fraud, as illustrated through data-driven Tables 1,2,3 and insightful Chart 1.

The chart below illustrates the impact of various Ind AS standards on key areas where financial fraud is commonly found.

Revenue Manipulation and Asset Overvaluation are the most frequent fraud risks addressed by Ind AS.

3.1. Major areas where Ind AS prevents fraud

The following pie chart in Graph 1 represents the percentage contribution of various Ind AS standards in preventing different types of financial fraud [37].

The chart highlights how IND AS addresses hidden liabilities (20 %) to ensure the company's financial health. It significantly curtails revenue manipulation (30 %) and asset overvaluation (25 %) through strict policies, such as fair value measurement and revenue recognition. Related-party transactions (15 %) are also addressed through mandatory disclosures, ensuring transparency. Other fraudulent activities (10 %) are minimised via a robust framework. IND AS enhances accountability, reliability, and investor confidence in financial statements.

4. Case study

Providing the full details of these case studies and their original balance sheets is restricted due to confidentiality and legal considerations, as corporate financial reports and internal documents may not be publicly available. However, this can be summarise each case with publicly available information and provide critical lessons from their financial reports in Table 4,5,6–53.

Table 1
Key Ind AS standards and their role in preventing financial frauds.

Ind AS Standard	Area Covered	Role in Preventing Fraud	Case Study Example
Ind AS 1	Presentation of Financial Statements	Ensures proper classification and disclosure, reducing the risk of manipulation and misstatement.	Case Study: IL&FS (Infrastructure Leasing & Financial Services): Manipulation in the classification of liabilities caused its financial collapse, drawing attention to proper financial reporting as per Ind AS 1 to avoid such incidents [26].
Ind AS 115	Revenue from Contracts with Customers	Prevents premature recognition of revenue, ensuring accurate reporting of earnings.	Case Study: Wipro - They applied Ind AS 115 for revenue recognition, demonstrating how proper standards prevent early revenue reporting and contribute to transparency [27].
Ind AS 109	Financial Instruments	Ensures proper classification, valuation, and disclosure of financial assets, reducing the risk of hidden liabilities.	Case Study: Yes Bank: The correct implementation of Ind AS 109 in evaluating financial instruments helped identify risk, especially related to NPAs (Non-Performing Assets) [28].
Ind AS 110	Consolidated Financial Statements	Mandates the consolidation of subsidiaries, preventing off-balance sheet liabilities and financial misrepresentation.	Case Study: Tata Sons: Ind AS 110 ensured proper consolidation, preventing off-balance sheet fraud involving subsidiaries [29].
Ind AS 36	Impairment of Assets	Prevents overvaluation of assets by requiring impairment tests to reflect an accurate financial position.	Case Study: Reliance Communications: Applied Ind AS 36 to adjust for impairment losses when the telecom sector experienced significant stress [30].
Ind AS 24	Related Party Disclosures	Requires transparent disclosure of transactions with related parties, reducing the risk of hidden financial fraud.	Case Study: Jet Airways: Failure to disclose related party transactions and financial misreporting could have been prevented by strict adherence to Ind AS 24 [31].
Ind AS 37	Provisions, Contingent Liabilities, and Assets	Ensures proper reporting of provisions and contingencies, avoiding underreporting of risks.	Case Study: Vodafone Idea: Adopted Ind AS 37 to disclose and manage contingent liabilities related to the AGR dues dispute with the government [32].
Ind AS 16	Property, Plant, and Equipment	Prevents overstatement of assets by requiring accurate depreciation and valuation methods.	Case Study: SAIL (Steel Authority of India): Followed Ind AS 16 to properly account for asset depreciation in heavy infrastructure investments [33].

Source: Authors' Research on Ind AS standards in financial fraud prevention: Case studies and analysis.

Table 2
Impact of Ind AS on Financial Fraud Prevention.

Aspect	Ind AS Involved	Impact on Fraud Prevention
Revenue Manipulation	Ind AS 115	High
Asset Overvaluation	Ind AS 36, Ind AS 16	High
Hidden Liabilities	Ind AS 109, Ind AS 110, Ind AS 37	High
Related Party Transactions	Ind AS 24	Medium
Premature Revenue Recognition	Ind AS 115	High
Financial Instruments	Ind AS 109	High

Source: Author's Compiled.

Table 3
Benefits of Ind AS in Enhancing Financial Reporting Integrity.

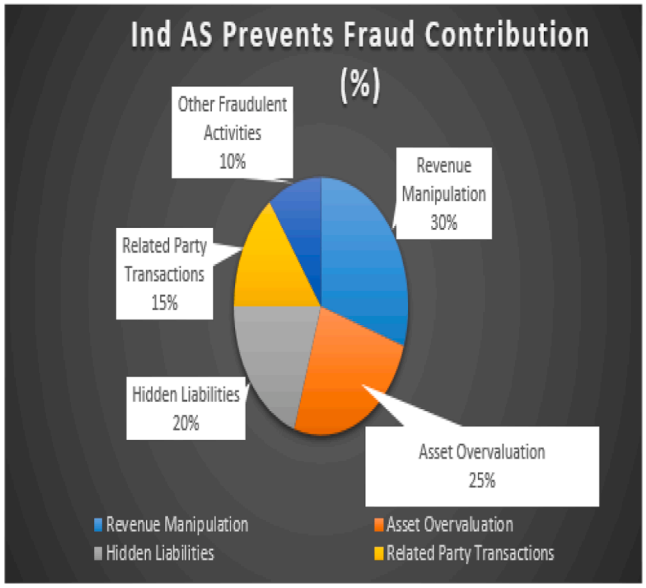
Benefit	Ind AS Feature	Impact	Company Example
Transparency	Comprehensive disclosure requirements	Reduces opportunities for fraud by identifying and addressing hidden or misclassified data.	Infosys: Known for its transparent disclosures, following Ind AS to maintain high financial integrity [34].
Fair Valuation	Emphasis on fair value measurement (Ind AS 109)	Limits on overstatement or understatement of assets and liabilities.	Yes Bank: Utilised fair value measurement under Ind AS 109 to expose risks in its asset portfolio [28].
Comparability	Consistent reporting across entities (Ind AS 1)	Reduces manipulation by enforcing uniform accounting practices.	Wipro: Implemented consistent reporting, providing clarity in financial statements [35].
Consolidation	Consolidation of financial statements (Ind AS 110)	Prevents off-balance sheet financing and undisclosed liabilities.	Tata Group: Effective consolidation of subsidiaries to prevent fraud [29].
Accurate Revenue Recognition	Stringent revenue recognition (Ind AS 115)	It prevents companies from inflating earnings through early revenue reporting.	Hindustan Unilever Limited (HUL): Followed strict revenue recognition policies and maintained reliable earnings reports [36].

Source: Authors' Research on Ind AS standards in financial fraud prevention: Case studies and analysis.

4.1. IL&FS (Infrastructure leasing & financial services) - manipulation of liabilities (Ind AS 1)

The case of Infrastructure Leasing & Financial Services (IL&FS) illustrates the abuse of accounting standards, particularly Ind AS 1, to misrepresent financial health. Analysis of IL&FS's reports from 2017 to 2024 shows a peculiar pattern in the Debt-to-Equity Ratio, which stayed at 0 %. This is unusual for infrastructure and financial firms, which typically depend on borrowed capital. Reporting a zero Debt-to-Equity Ratio raises concerns about concealing off-balance sheet liabilities, creating a misleading financial image. This exemplifies financial window dressing, as liabilities may have been manipulated to mislead stakeholders, investors, and regulators about the actual financial condition.

IL&FS's financial ratio analysis suggests potential manipulation. The



Graph 1. Pie Chart: Major Areas Where Ind AS Prevents Fraud.
Source: Author's Compiled.

Current Ratio surged from 7.4 % in 2017 to 318.2 % in 2024, indicating inflated current assets or understated liabilities. The Quick Ratio reflected this trend. The Return on Equity (ROE) dropped sharply in 2018 and fluctuated, while the Debt-to-Equity Ratio remained at 0 %, raising further concerns. Although Ind AS aims to enhance transparency, the IL&FS case shows that even strong standards can be misused without strict enforcement.

The rules used for triggering the "Fraud Suspected" indicator are based on significant anomalies in Return on Equity (ROE) and liquidity ratios. Rule 1 is activated when ROE drops by >70 % in a single year, which may suggest an abrupt decline in performance. Rule 2 flags concern when the ROE becomes negative, accompanied by a drop of >30 % in liquidity, indicating possible prolonged financial stress or manipulation. Rule 3 is triggered when ROE falls by over 500 %, marking an extreme anomaly and a strong potential red flag.

These rules, applied to the financial data, identified suspicions of fraud in 2018 (Rule 1), 2021 (Rule 2), and 2022 (Rule 3). Although these do not confirm fraudulent activity, such drastic shifts in performance and liquidity warrant a deeper forensic investigation to rule out misrepresentation or hidden operational issues.

4.1.1. *Corporate governance adoption*

IL&FS (Infrastructure Leasing & Financial Services) faced a corporate governance crisis due to liability manipulation, breaching Ind AS 1. A substantial decline in return on equity (ROE) in 2018, combined with increased liquidity issues, indicated performance challenges and heightened governance scrutiny. Although some governance improvements were made, they were reactive rather than proactive. Efforts were inconsistent, with no significant reforms in 2019 and 2020. More decisive actions emerged in 2021 and 2022, aligning with financial recovery. However, governance initiatives weakened again in 2023 and 2024 despite improved performance. This inconsistency highlights IL&FS's failure to implement strong governance practices, relying on temporary solutions during crises instead of fostering long-term accountability.

Table 5
Suspected fraud analysis.

Year	ΔROE	ΔCR	ΔQR	Fraud Suspected	Rule Triggered
2017	–	–	–	0	–
2018	(−0.747)	(8.243)	(8.243)	1	Rule 1 – Sharp ROE fall
2019	(1.049)	(−0.088)	(−0.088)	0	–
2020	(−0.202)	(0.258)	(0.258)	0	–
2021	(−1.433)	(−0.381)	(−0.381)	1	Rule 2 – ROE negative + poor liquidity
2022	(−5.276)	(0.235)	(0.235)	1	Rule 3 – Extremely poor ROE
2023	(0.718)	(1.187)	(1.187)	0	–
2024	(−0.070)	(1.425)	(1.425)	0	–

Source: Author's computation.

Table 6
Method of suspected fraud indicator (Adapted for ROE & Liquidity).

Indicator	Calculation	Interpretation Logic
ΔROE (%)	$ROE_t - ROE_{t-1}$	Large drops may indicate operational distress or earnings management
ΔCR (%) / ΔQR (%)	$CR_t - CR_{t-1}$ / $QR_t - QR_{t-1}$	Sharp decreases alongside negative ROE can suggest liquidity stress
Fraud Suspected	Binary (0 = No, 1 = Yes)	1 is assigned when any rule is triggered

Source: Author compiled.

Table 7 shows that the descriptive statistics indicate significant data asymmetry and a non-normal distribution. The average ROE is 10.75, characterised by low variability yet high skewness (8.26), which suggests the presence of extreme positive outliers. The Current and Quick Ratios, with a mean of 96.84 and a standard deviation of 33.86, also demonstrate high skewness and leptokurtic patterns, reflecting a concentration of high values. The adoption of Indas is prevalent, with a mean of 0.88, accompanied by extreme kurtosis. The variables of pandemic, Suspected Fraud, and ECG exhibit clustering at lower values, implying modelling difficulties due to their skewed and peaked distributions.

The one-sample *t*-test results in Table 8 indicate that Return on Equity (ROE), Current Ratio, and Quick Ratio are significantly higher than the benchmark value of 2.365, with p-values of 0.024 (ROE), 0.027 (Current Ratio), and 0.027 (Quick Ratio), respectively. This indicates that these financial metrics exceed the benchmark statistically. Conversely, IndAS status, Pandemic, Suspected Fraud, and Enhanced Corporate Governance show significant negative t-values ($p < 0.001$), indicating their means fall below the benchmark value. Additionally,

Table 4
Comparative analysis of IL&FS (infrastructure leasing & financial services) (2017–2024).

Year	Debt/Equity	ROE	CR	QR	IndAS	Pandemic	SF	Enhance Corporate Governance
2017	0	16.2	7.4	7.4	0	0	0	0
2018	0	4.1	68.4	68.4	1	0	1	1
2019	0	8.4	62.4	62.4	1	0	0	0
2020	0	6.7	78.5	78.5	1	1	0	0
2021	0	−2.9	48.6	48.6	1	1	1	1
2022	0	12.4	60.0	60.0	1	1	1	1
2023	0	21.3	131.2	131.2	1	0	0	0
2024	0	19.8	318.2	318.2	1	0	0	0

Source: Author's calculations with company annual reports (https://www.ilfsindia.com/).

Table 7
Descriptive analysis of IL&FS (Infrastructure Leasing & Financial Services).

Descriptive Statistics								
	N	Mean		Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Std. Error	Statistic	Statistic	Std. Error	Statistic	Std. Error
ROE	8	10.75	2.92	8.26	−0.28	0.75	−0.71	1.48
Current Ratio	8	96.84	33.86	95.76	2.16	0.75	5.19	1.48
Quick Ratio	8	96.84	33.86	95.76	2.16	0.75	5.19	1.48
IndAS status	8	0.88	0.13	0.35	−2.83	0.75	8.00	1.48
Pandemic	8	0.38	0.18	0.52	0.64	0.75	−2.24	1.48
Suspected Fraud	8	0.38	0.18	0.52	0.64	0.75	−2.24	1.48
Enhance Corporate Governance	8	0.38	0.18	0.52	0.64	0.75	−2.24	1.48
Valid N (list-wise)	8							

Source: Through SPSS compiled by the Author.

Table 8
One-sample test.

One-Sample Test						
	Test Value = 2.365					
	t	df	Sig. (2-tailed)	Mean Difference	95 % Confidence Interval of the Difference	
					Lower	Upper
ROE	2.87	7	.024	8.38	1.47	15.29
Current_Ratio	2.79	7	.027	94.47	14.41	174.53
Quick_Ratio	2.79	7	.027	94.47	14.41	174.53
IndAS_status	−11.92	7	.000	−1.49	−1.79	−1.19
Pandemic	−10.87	7	.000	−1.99	−2.42	−1.56
Suspected_Fraud	−10.87	7	.000	−1.99	−2.42	−1.56
Enhance_Corporate_Governance	−10.87	7	.000	−1.99	−2.42	−1.56

Source: Through SPSS compiled by the Author.

narrow confidence intervals reinforce these crucial differences. These findings suggest a substantial divergence from the benchmark, highlighting potential structural or policy-related disparities within the dataset.

The multinomial logit results in Table 9 indicate that Return on Equity (ROE) harms IndAS adoption (β = –1.5587, p = 0.0000). In contrast, the Current Ratio (CR) has a positive influence on it (β = 0.6508, p = 0.0000). This suggests that companies with strong liquidity and lower profitability are more inclined to adopt Indas. However, both ROE and CR do not significantly impact the outcomes related to the Pandemic, Suspected Fraud (SF), or Enhanced Corporate Governance (ECG), indicating their limited predictive capability in these areas.

Table 10 summarises model fit statistics for four models. Model 1 (IndAS) stands out with perfect 100 % correct predictions and a statistically significant likelihood ratio test (p = 0.0039), indicating an excellent fit. In contrast, Models 2, 3, and 4 show lower correct predictions (75 %) and non-significant likelihood ratio tests, suggesting a poorer overall fit. This highlights Model 1′s superior Performance in explaining the observed data.

4.2. Wipro- revenue recognition (Ind AS 115)

The Company was established on April 13, 2015, and has been implementing Ind AS since then. Accounting policies have been consistently applied in these financial statements. Wipro Limited, based on the same date, adopted Ind AS 115 - Revenue from Contracts with Customers, promoting uniform revenue recognition from the outset. Ind AS 115 emphasises revenue recognition based on control transfer, improving transparency and comparability.

Table 9
Combined Multinomial Logit Model Estimation Results (2017–2024).

Variable	Model 1: IndAS (β, p-value)	Model 2: Pandemic (β, p-value)	Model 3: SF (β, p-value)	Model 4: ECG (β, p-value)
ROE	–1.5587, 0.0000	–0.1030, 0.3002	–0.1308, 0.2895	–0.1308, 0.3157
CR	0.6508, 0.0000	0.0013, 0.8661	0.0029, 0.7374	0.0029, 0.8123

Source: Author's Compiled.

Table 11 illustrates that Wipro's revenue growth fluctuated between 2017 and 2024, with positive increases observed in 2019, 2020, 2022, and 2023, while declines were noted in 2018 and 2024. The gross profit margin varied from 21.1 % to 27.7 %, and the net profit margin ranged from 12.4 % to 17.5 %, reflecting effective cost control. The return on equity reached a high of 19.7 % in 2021 but decreased to 14.8 % by 2024. Following Ind AS 115, Wipro's profitability demonstrates a dedication to transparent reporting and sound financial management [38].

We will apply three rules to detect suspected fraud, focusing on key financial metrics outlined in Table 12. Rule 1 flags a substantial positive spike in Revenue Growth Rate (change > 10 %) from the previous year, alongside significant drops in Gross Profit Margin (change < −2 %) and Net Profit Margin (change < −2 %) in the current year. Rule 2 flags a notable decrease in Gross Profit Margin (change < −5 %). Lastly, Rule 3 indicates concern for a considerable drop in Net Profit Margin (change < −5 %). By applying these rules to the financial data, we can evaluate potential indicators of fraud over time.

4.2.1. Corporate governance adoption: 2015–2024

From 2015 to 2016, Wipro made minimal progress in its corporate governance practices, continuing to adhere to Indian Generally Accepted Accounting Principles (GAAP) while preparing to transition to International Financial Reporting Standards (IFRS) or Ind AS. During the 2017–2018 period, Wipro began adopting Ind AS, resulting in improved financial transparency and governance, particularly with the introduction of Ind AS 115 for revenue recognition. Starting in 2019, Wipro has strengthened its governance framework through enhanced board

Table 10
Model fit statistics summary.

Model	Log-Likelihood	AIC	BIC	HQ	Correct Predictions	Likelihood Ratio Test (χ^2, p-value)
Model 1: IndAS	–0.00000000406	4.000	4.159	2.928	8 / 8 (100 %)	$\chi^2(4) = 11.090$, $p = 0.0039$
Model 2: Pandemic	–4.4153	12.831	12.990	11.759	6 / 8 (75 %)	$\chi^2(4) = 2.260$, $p = 0.3231$
Model 3: SF	–4.1335	12.267	12.426	11.195	6 / 8 (75 %)	$\chi^2(4) = 2.823$, $p = 0.2437$
Model 4: ECG	–4.1335	12.267	12.426	11.195	6 / 8 (75 %)	$\chi^2(4) = 2.823$, $p = 0.2437$

Source: Author's Compiled.

Table 11
Comparative analysis (2017–2024).

Year	Revenue Growth Rate (%)	Gross Profit Margin (%)	Net Profit Margin (%)	ROE (%)	IndAS	Pandemic	SF	ECG
2015	4.1	27.1	18.5	23.3	0	0	0	0
2016	2.6	26.4	17.5	19.3	0	0	0	0
2017	2.7	24.1	15.4	16.4	0	0	0	0
2018	−3.1	23.7	14.7	16.7	1	0	0	1
2019	7.4	25.1	15.3	15.9	1	0	0	1
2020	4.3	24.6	15.9	17.6	1	1	0	1
2021	−0.4	27.7	17.5	19.7	1	1	0	1
2022	22	23.6	15.4	18.7	1	1	1	1
2023	9.1	21.1	12.6	14.6	1	0	0	1
2024	−0.4	21.6	12.4	14.8	1	0	0	1

Source: Author's calculations with company annual reports (https://www.wipro.com).

Table 12
Suspected fraud analysis.

Year	Revenue Growth Rate (%)	Gross Profit Margin (%)	Net Profit Margin (%)	ROE (%)	Revenue Growth Rate Change (YoY)	Gross Profit Margin Change (YoY)	Net Profit Margin Change (YoY)	Fraud_Suspected
2015	4.1	27.1	18.5	23.3	N/A	N/A	N/A	0
2016	2.6	26.4	17.5	19.3	−1.5	−0.7	−1	0
2017	2.7	24.1	15.4	16.4	0.1	−2.3	−2.1	0
2018	−3.1	23.7	14.7	16.7	−5.8	−0.4	−0.7	0
2019	7.4	25.1	15.3	15.9	10.5	1.4	0.6	0
2020	4.3	24.6	15.9	17.6	−3.1	−0.5	0.6	0
2021	−0.4	27.7	17.5	19.7	−4.7	3.1	1.6	0
2022	22	23.6	15.4	18.7	22.4	−4.1	−2.1	1
2023	9.1	21.1	12.6	14.6	−12.9	−2.5	−2.8	0
2024	−0.4	21.6	12.4	14.8	−9.5	0.5	−0.2	0

Source: Author's Compiled.

oversight, expanded ESG initiatives, increased ethical disclosures, and improved risk management. By 2023 and 2024, the company had aligned with global governance standards, providing integrated reports and maintaining high levels of transparency and accountability.

Table 13 presents statistics for financial and governance variables over a decade, highlighting key trends. The average revenue growth rate is 4.83 % with a standard deviation of 7.05, indicating variability. The distribution has skewness of 1.72 and kurtosis of 3.85, classifying it as right-skewed and leptokurtic. The gross and net profit margins average 24.50 % and 15.52 %, respectively, with slight negative skewness and low kurtosis, reflecting mild asymmetry. Return on Equity (RoE) averages 17.70 % with a positive skew of 0.94, indicating occasional high values. The IndAS, Pandemic, and Enhance Corporate Governance variables are binary, with means of 0.70, 0.30, and 0.70. The Suspected Fraud variable has a low mean of 0.10 but is highly skewed, showing rare but extreme occurrences. These statistics provide insights into financial stability and governance practices.

The one-sample *t*-test results in Table 14 reveal significant findings.

Table 13
Descriptive statistics.

Descriptive Statistics								
	N	Mean		Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Std. Error	Statistic	Statistic	Std. Error	Statistic	Std. Error
Revenue Growth Rate	10	4.83	2.23	7.05	1.72	.68	3.85	1.33
Gross Profit Margin	10	24.50	.68	2.17	−0.10	.68	−0.73	1.33
Net Profit Margin	10	15.52	.63	1.99	−0.24	.68	−0.52	1.33
RoE	10	17.70	.83	2.64	.94	.68	.95	1.33
IndAS	10	.70	.15	.48	−1.03	.68	−1.22	1.33
Pandemic	10	.30	.15	.48	1.03	.68	−1.22	1.33
Suspected Fraud	10	.10	.10	.31	3.16	.68	10.00	1.33
Enhance Corporate_Governance	10	.70	.15	.48	−1.03	.68	−1.22	1.33
Valid N (listwise)	10							

Source: Through SPSS compiled by the Author.

With a test value of 2.262, Gross Profit Margin, Net Profit Margin, and Return on Equity (RoE) exceed this value significantly, with low p-values ($p = 0.000$) against the null hypothesis. Their mean differences are significant, and the 95 % confidence intervals do not include zero, confirming statistical significance. In contrast, the Revenue Growth Rate ($t = 1.152$, $p = 0.279$) does not differ significantly from the test value, as indicated by a wide confidence interval that encompasses both negative and positive values. IndAS, Pandemic, Suspected Fraud, and Enhance Corporate Governance exhibit significant negative t-values and p-values ($p = 0.000$), indicating their mean occurrences are lower than the benchmark. Overall, findings indicate strong financial performance in profitability metrics, while governance-related variables show consistent patterns throughout the assessed period.

Table 15 presents the estimation results from the Combined Multinomial Logit Model for the period from 2015 to 2024. It highlights the effects of key financial variables on various outcome models: ECG, IndAS, SF, and Pandemic Impact. The Revenue Growth Ratio, ROE, Gross Profit Margin, and Net Profit Margin are statistically significant ($p < 0.01$) in Models 1 (ECG), 2 (IndAS), and 3 (SF), indicating they strongly predict these outcomes. Notably, "Net

Profit Margin" shows a significant adverse effect in Models 1 and 2, while "Gross Profit Margin" shows a significant positive effect. In contrast, these financial variables reveal no statistical significance (NS) in Model 4 (Pandemic Impact), indicating they do not predict pandemic-related outcomes. This highlights the varying predictive power of financial metrics, depending on the specific model analysed.

When assessing statistical models in the Table 16, several criteria evaluate their fit and predictive ability. AIC (Akaike information criterion) the model adapts too much to training data, leading to poor generalizability. Lastly, the Likelihood Ratio Test evaluates the overall significance of predictors; a statistically significant p-value (generally below 0.05) indicates strong explanatory power

4.3. Yes, bank - financial instruments (Ind AS 109)

The introduction of Indian Accounting Standard (Ind AS) 109 – Financial Instruments brought about a significant change in how banking institutions, such as Yes Bank, report financials. This change primarily affects the classification, measurement, and impairment of financial assets and liabilities. From 2017 to 2024, Yes Bank has faced various challenges and recoveries related to asset quality, risk management, and profitability, all of which have been shaped by the

Table 14
One-Sample Test.

One-Sample Test						
	Test Value = 2.262					
	t	df	Sig. (2-tailed)	Mean Difference	95 % Confidence Interval of the Difference	
					Lower	Upper
Revenue Growth Rate	1.152	9	.279	2.56800	−2.4766	7.6126
Gross Profit Margin	32.338	9	.000	22.23800	20.6824	23.7936
Net Profit Margin	20.999	9	.000	13.25800	11.8298	14.6862
RoE	18.470	9	.000	15.43800	13.5471	17.3289
IndAS	−10.226	9	.000	−1.562	−1.91	−1.22
Pandemic	−12.844	9	.000	−1.962	−2.31	−1.62
Suspected Fraud	−21.620	9	.000	−2.162	−2.39	−1.94
Enhance Corporate Governance	−10.226	9	.000	−1.562	−1.91	−1.22

Source: Through SPSS compiled by the Author.

Table 15
Combined multinomial logit model estimation results (2015–2024).

Variable	Model 1: ECG (β, p-value)	Model 2: IndAS (β, p-value)	Model 3: SF (β, p-value)	Model 4: Pandemic Impact (β, p-value)
Revenue Growth Ratio	19.7836, 0.0000	19.7836, 0.0000	2.4959, 0.0000	0.1153, 0.2657
ROE (Return on Equity)	62.7675, 0.0000	62.7675, 0.0000	−3.3416, 0.0000	−0.2876, 0.6102
Gross Profit Margin	130.0130, 0.0000	130.0130, 0.0000	−9.7285, 0.0000	−0.6073, 0.2836
Net Profit Margin	−274.8960, 0.0000	−274.8960, 0.0000	16.7503, 0.0000	1.1897, 0.3668

Source: Author's Compiled.

requirements of Ind AS 109.

In Table 17, prior to its asset quality crisis, Yes Bank had recorded low gross NPA levels. However, following the implementation of Ind AS 109, NPAs skyrocketed to 16.80 % in 2020, highlighting previously obscured risks. Recovery efforts gradually decreased non-performing assets (NPAs) to 1.70 % by 2024. The Provision Coverage Ratio declined during the crisis but rebounded to 428 %, and the Return on Assets (RoA) also improved. Ind AS 109 increased transparency, uncovering vulnerabilities and aiding long-term recovery alongside enhanced risk management.

Table 18 shows a rise in GNPA and NNPA in 2019–2020, a sharp decline in PCR and RoA, and a decrease in CAR, which triggered fraud detection in 2020. Fraud reappeared in 2021 and 2023 with varying rules, indicating financial distress patterns linked to NPA trends and poor asset quality.

Table 19 outlines a method to detect potential fraud using key financial indicators. Variations in Gross and Net Non-Performing Assets (NPAs) indicate asset quality and credit risk. A decreasing Provision Coverage Ratio (PCR) alongside rising Non-Performing Assets (NPAs) suggests inadequate provisioning. Consecutive years of negative Return on Assets (RoA) reflect operational strain. A Capital Adequacy Ratio (CAR) dropping below 10 % signals a weak capital buffer. Visual indicators highlight the severity of these changes. Fraud suspicion is marked as binary (1 = Yes) based on significant GNPA increases with declining PCR, sustained negative RoA, or dubious provisioning adjustments.

4.3.1. Corporate governance adoption

From 2017 to 2018, Yes Bank projected effective governance, concealing serious issues like promoter control and inadequate risk management. By 2019, governance flaws emerged due to concerns from the RBI and the departure of CEO Rana Kapoor amid allegations of fraud. The pivotal year of 2020 saw the RBI initiate restructuring, reorganise leadership, and address shortcomings. Kapoor's arrest unveiled further misconduct. From 2021 to 2024, the bank focused on regaining trust through board reforms, policy enforcement, and increased transparency, highlighting efforts to restore strong corporate governance and enhance credibility.

Table 20 presents descriptive statistics for eight variables over eight years, highlighting significant insights. Both Gross NPA and Net NPA exhibit moderate positive skewness, indicating a right-tailed distribution, whereas CAR displays strong negative skewness, suggesting a left-tailed concentration. CAR's high kurtosis (6.169) indicates sharp peakedness, while most variables exhibit negative kurtosis, reflecting flatter distributions. RoA has the highest variability with a standard deviation of 2.94, indicating fluctuating profitability. Binary variables, such as pandemic, suspected fraud, and governance enhancement,

Table 16
Model fit statistics summary.

Model	Log-Likelihood	AIC	BIC	HQ	Correct Predictions	Likelihood Ratio Test (χ^2, p-value)
Model 20 (ECG)	−7.57e−08	8.0000	9.2103	6.6723	10 / 10 (100 %)	$\chi^2(4) = 13.863$, $p = 0.0077$
Model 21 (IndAS)	−7.57e−08	8.0000	9.2103	6.6723	10 / 10 (100 %)	$\chi^2(4) = 13.863$, $p = 0.0077$
Model 15 (SF)	−4.25e−09	8.0000	9.2103	6.6723	10 / 10 (100 %)	$\chi^2(4) = 13.863$, $p = 0.0077$
Model 19 (Pandemic)	−5.3721	18.7442	19.9546	17.4165	8 / 10 (80 %)	$\chi^2(4) = 3.1187$, $p = 0.5382$

Source: Author's Compiled.

Table 17
Analysis of Yes bank - financial instruments (Ind AS 109) (2017–2024).

year	GNPA	NNPA	PCR	RoA	CAR	Pandemic	IndAS	Suspected Fraud	Enhance Corporate Governance
2017	1.52	0.81	46.88	1.8	17	0	0	0	0
2018	1.28	0.64	50.02	1.6	18	0	0	0	0
2019	3.22	1.86	43.1	0.5	17	0	1	0	0
2020	16.8	5.03	73.77	−5.1	8	1	1	1	1
2021	15.41	5.88	65.7	−5.7	17	1	1	1	1
2022	13.93	4.532	70.67	0.4	17	1	1	0	1
2023	2.17	0.83	62.27	0.2	18	0	1	1	1
2024	1.7	0.6	66.61	0.3	15	0	1	0	1

Source: Author's calculations with company annual reports(https://www.yesbank.in).

Table 18
Suspected fraud analysis from NPA & financial ratios.

Year	ΔG-NPA	ΔN—NPA	PCR	RoA	CAR	Fraud_Suspected_	Rule Triggered
2017	–	–	571	1.80	17	0	–
2018	−0.24	−0.17	421	1.60	18	0	–
2019	1.9	1.22	227	0.50	17	0	–
2020	13.58	3.17	52	−5.10	8	**1**	**Rule 1 + 2**
2021	−1.39	0.85	47	−5.70	17	**1**	**Rule 2**
2022	−1.48	−1.35	54	0.40	17	0	–
2023	−11.76	−3.7	434	0.20	18	**1**	**Rule 3**
2024	−0.47	−0.23	428	0.30	15	0	–

Source: Author's Compiled.

exhibit low means and identical statistics due to limited variation. Overall, the data indicates a non-normal distribution with some variables exhibiting skewed behaviour and differing dispersion trends.

Table 21 shows the one-sample *t*-test results for Yes Bank's indicators with a test value of 2.365. The Gross NPA and Net NPA do not differ significantly from the benchmark ($p > 0.05$), indicating no substantial deviation from the benchmark. However, the Provision Coverage Ratio ($p = 0.000$) and Capital Adequacy Ratio (CAR) ($p = 0.000$) exhibit significant increases, highlighting substantial capital and provisioning buffers. The Return on Assets (RoA) is notably lower ($p = 0.020$), indicating weak profitability. Factors such as the pandemic, IndAS, suspected fraud, and governance enhancement show significant negative mean differences ($p = 0.000$), reflecting operational or regulatory challenges faced during this period.

The results from the four models presented in Table 22—ECG, IndAS, SF, and Pandemic Impact—demonstrate differing influences of financial indicators. GNPA has a significant impact on all models, causing adverse effects in ECG and IndAS, while positively influencing SF. NNPA is insignificant in ECG but shows high significance in the other models, with positive effects in IndAS and negative ones in SF and Pandemic Impact. PCR consistently proves significant across the models, positively impacting ECG and IndAS, although it has a negative influence on SF and Pandemic Impact. RoA and CAR are mainly significant, with RoA remaining consistently negative. CAR shows varied effects, resulting in significantly adverse outcomes in ECG and IndAS but positive results in Suspected fraud.

Table 23 illustrates that all four models—Pandemic Impact, IndAS, SF, and ECG—achieve perfect predictive accuracy (100 %) with matching log-likelihood, AIC, BIC, and HQ values. The results from the Likelihood Ratio Test are statistically significant for all models ($\chi^2 = 11.090$, $p = 0.0496$), indicating that each model effectively accounts for the variation in the dependent variable.

4.4. Tata sons - consolidation (Ind AS 110)

Ind AS 110 outlines that the Independent Auditor's Report on Tata International Limited's Consolidated Financial Statements for the year ended March 31, 2020, evaluates internal financial controls by Section 143(3)(i) of the Companies Act, 2013. Auditors assessed these controls for the Holding Company and its subsidiaries, associates, and joint ventures in India. Ensuring adequate internal controls falls to the Boards of Directors, which is crucial for maintaining the effectiveness of financial processes, protecting assets, detecting fraud, and ensuring accurate reporting. The auditor's role was to express an opinion on the organisation's internal controls and financial statements. In 2017, Tata Sons changed from a public limited to a private limited company, contested by former executive chairman Cyrus Mistry. In 2019, the NCLAT deemed this conversion and Chairman Chandrasekaran's appointment illegal, reinstating Mistry.

Table 24 shows that between 2018 and 2024, Tata Sons underwent significant changes in its financial and governance metrics. The Debt/Equity Ratio sharply increased in 2020 due to the pandemic, signalling financial stress, while the Growth Assets Ratio fluctuated and became negative during 2020–2021. The adoption of Ind AS in 2018 enhanced transparency. While allegations of fraud emerged during the pandemic years, governance practices consistently improved after 2020, indicating a substantial evolution in corporate governance from 2018 onward. These trends reflect strategic adjustments and resilience in the face of economic difficulties and regulatory shifts, thereby reinforcing long-term corporate stability.

Table 19
Method of suspected fraud indicator.

Indicator	Calculation Method	Interpretation Logic
ΔG-NPA (%)	G-NPAt−*G*-NPAt−1	A significant increase may signal a deterioration in asset quality.
ΔN—NPA (%)	N-NPAt−*N*—NPAt−1	An increase indicates ineffective provisioning and rising credit risk.
PCR (%)	Reported directly, trend analysis is used	Falling PCR with rising NPAs suggests inadequate provisioning.
RoA (%)	Reported directly	A negative Return on Assets (RoA) over multiple years signals poor profitability or potential manipulation.
CAR (%)	Reported directly	A fall below 10 % is considered a sign of a weak capital cushion.
Indicators (Increase (I) /Decrease (D))	Based on thresholds: ±0.1 to ±1.0 = *I* or D; ±1.0 to ±5.0 = more I or More D; > ±5.0 = High I/ High D	Used for visual representation of the intensity of change in financial indicators.
Fraud Suspected	Binary (0 = No, 1 = Yes), based on rule trigger	1 is assigned when specific patterns or rules indicating manipulation or risk are detected.
Rule 1	ΔG—NPA > 5 % and PCR falls sharply	Indicates deterioration in asset quality due to inadequate provisioning.
Rule 2	RoA < 0 for two or more consecutive years	Suggests ongoing operational or financial stress.
Rule 3	G-NPA drops drastically, and PCR increases sharply	May indicate possible window dressing or manipulation in bad loan reporting or provisioning.

Source: Author's Compiled.

Table 25 presents the suspected fraud analysis for Tata Sons, revealing anomalies in 2020 and 2021. In 2020, the debt-to-equity ratio saw a significant increase, triggering Rule 1 and suggesting potential financial distress. In 2021, a dramatic decrease in assets triggered Rule 2. These two years raised flags for suspected fraud, whereas the other years showed stability with no rule triggers or indications of fraud.

4.4.1. Corporate governance adoption

Tata Sons' corporate governance has evolved since 2018, following the implementation of Indas, marked by increased transparency in debt and assets. By 2020, governance had gained importance due to a rising debt-to-equity ratio, the challenges posed by COVID-19, and allegations of fraud, which exposed systemic weaknesses and prompted enhanced oversight. Since 2021, Tata Sons has strengthened its governance commitment, achieving better financial stability and maintaining fraud-free operations. In summary, regulatory changes and crises have driven Tata Sons to adopt a stricter governance framework.

Table 21
Yes, Bank, One-Sample T-Test.

One-Sample Test						
	Test Value = 2.365					
	t	df	Sig. (2-tailed)	Mean Difference	95 % Confidence Interval of the Difference	
					Lower	Upper
Gross NPA	1.874	7	.103	4.63875	−1.2155	10.4930
Net NPA	.199	7	.848	.15775	−1.7136	2.0291
Provision Coverage Ratio	14.023	7	.000	57.51250	47.8141	67.2109
RoA	−3.001	7	.020	−3.11500	−5.5695	−0.6605
CAR	11.531	7	.000	13.51000	10.7395	16.2805
Pandemic	−10.875	7	.000	−1.990	−2.42	−1.56
IndAS	−9.868	7	.000	−1.615	−2.00	−1.23
Suspected fraud	−10.875	7	.000	−1.990	−2.42	−1.56
Enhance Corporate Governance	−9.509	7	.000	−1.740	−2.17	−1.31

Source: Through SPSS compiled by the Author.

The descriptive analysis of Tata Sons from 2015 to 2024, shown in Table 26, indicates a significant average Debt/Equity Ratio (mean = 5.898) with considerable variability, suggesting occasional financial difficulties. On average, asset growth reached 17.03 %, although there was some skewness in the data. The implementation of IndAs averaged 70 %, while the pandemic influenced 30 % of the examined years. Suspected fraud occurred in 20 % of the cases. Additionally, emerging corporate governance practices were noted in 60 % of instances, reflecting progressive improvements in governance over time. Refer to Table 26.

The one-sample *t*-test results for Tata Sons, shown in Table 27, indicate that the Debt/Equity Ratio and Growth in Assets are not significantly different from the test value of 2.262 ($p > 0.05$). Conversely, the adoption of Indas, the effects of the pandemic, allegations of fraud, and emerging corporate governance practices exhibit statistically significant differences ($p < 0.001$). These findings highlight significant deviations in accounting standards, the crisis's impact, indicators of fraud, and advancements in governance compared to expected norms during the analysed period

The multinomial logit analysis results for Tata Sons (2015–2024), presented in Table 28, reveal a significant correlation between a higher Debt/Equity Ratio and an increased likelihood of adopting IndAS ($\beta = 1.2598$, $p = 0.0060$), as well as enhancing corporate governance ($\beta = 1.5122$, $p = 0.0098$). This suggests a strategic shift toward enhanced

Table 20
Yes, Bank descriptive analysis.

Descriptive Statistics							
	N	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Gross NPA	8	7.0038	7.00254	.664	.752	−2.045	1.481
Net NPA	8	2.5228	2.23844	.641	.752	−1.812	1.481
Provision Coverage Ratio	8	59.8775	11.60063	−0.417	.752	−1.658	1.481
RoA	8	−0.7500	2.93598	−1.269	.752	−0.113	1.481
CAR	8	15.8750	3.31393	−2.429	.752	6.169	1.481
Pandemic	8	.38	.518	.644	.752	−2.240	1.481
IndAS	8	.75	.463	−1.440	.752	.000	1.481
Suspected fraud	8	.38	.518	.644	.752	−2.240	1.481
Enhance Corporate Governance	8	.63	.518	−0.644	.752	−2.240	1.481
Valid N (list-wise)	8						

Source: Through SPSS compiled by the Author.

Table 22
Combined multinomial logit model estimation results.

Variable	Model 1: ECG & lt;br>(beta, p-value)	Model 2: IndAS & lt;br>(beta, p-value)	Model 3: SF & lt;br>(beta, p-value)	Model 4: Pandemic Impact & lt;br>(beta, p-value)
GNPA	−7.8368, 0.0468	−22.3507, (0.0000)	34.9279, 0.0000	9.44520, 4.57e-11
NNPA	14.7203, 0.1953	64.7880, 0.0000	−142.9960, 0.0000	−17.5174, 1.63e-05
PCR	3.4067, 0.0000	2.35804, 0.0000	−2.6930, 0.0000	−0.499889, 0.0011
RoA	−1.7310, 0.1857	−15.6870, 0.0000	−48.6751, 0.0000	−1.13391, 0.0313
CAR	−10.4023, 0.0000	−7.12446, 0.0000	13.3269, 0.0000	0.248950, 0.6403

Source: Author's Compiled.

Table 23
Model Fit Statistics Summary.

Model	Log-Likelihood	AIC	BIC	HQ	Correct Predictions	Likelihood Ratio Test (chi2, p-value)
Model 1: Pandemic Impact	−6.47e-09	10.00000	10.39721	7.320994	8 / 8 (100 %)	chi2=11.0904,$p = 0.0496$
Model 2: IndAS	−3.74e-09	10.00000	10.39721	7.320994	8 / 8 (100 %)	chi2(5)=11.090,$p = 0.0496$
Model 3: SF	−7.88e-09	10.00000	10.39721	7.320994	8 / 8 (100 %)	chi2(5)=11.090,$p = 0.0496$
Model 4: ECG	−7.01e-09	10.00000	10.39721	7.320994	8 / 8 (100 %)	chi2(5)=11.090,$p = 0.0496$

Source: Author's Compiled.

Table 24
Tata Sons comparative analysis (2018–2024).

Year	Debt/Equity Ratio	Growth Assets Ratio	IndAS	Pandemic	Suspected Fraud	Emergence of Corporate Governance
2015	0.41	15.28	0	0	0	0
2016	0.37	1.64	0	0	0	0
2017	0.74	17.01	0	0	0	0
2018	1.76	91.23	1	0	0	**1**
2019	0.23	22.19	1	0	0	0
2020	48.51	−2.3	1	1	1	**1**
2021	1.63	−41.34	1	1	1	**1**
2022	1.66	30.86	1	1	0	**1**
2023	1.79	10.49	1	0	0	**1**
2024	1.88	25.3	1	0	0	**1**

Source: Author's calculations based on Tata Sons annual reports (tata.com, About Us).

transparency in the context of financial leverage. In contrast, the Growth Assets Rate has a notably adverse effect on the likelihood of fraudulent activities ($\beta = –10.6839$, $p < 0.0001$), indicating that a decrease in asset growth greatly heightens the risk of fraud. The other variables do not exhibit significant effects, underscoring the predominant role of financial structure and performance in important governance and fraud-related outcomes.

In Table 29, Model 3 (Suspected Fraud) exhibits the best fit, with 100 % prediction accuracy, the lowest AIC/BIC values, and a highly significant chi-square ($p = 0.0010$). Other models (IndAS, Pandemic, ECG) have moderate fits, with 70–90 % accuracy and marginal statistical significance.

4.5. Reliance communications - impairment losses (Ind AS 36)

The implementation of Ind AS 36 — Impairment of Assets — has fundamentally transformed financial reporting practices among Indian companies, especially those in financial distress, such as Reliance

Table 25
Suspected fraud analysis of Tata Sons.

Year	Debt/Equity Ratio	Δ D/E Ratio	Asset Growth %	Rule 1 ($\Delta D/E > 10$)	Rule 2 (Asset Growth < -30 %)	Rule 3 ($D/E > 10$ & Neg. Growth)	Fraud Suspected
2015	0.41	–	15.28	0	0	0	0
2016	0.37	−0.04	1.64	0	0	0	0
2017	0.74	0.37	17.01	0	0	0	0
2018	1.76	1.02	91.23	0	0	0	0
2019	0.23	−1.53	22.19	0	0	0	0
2020	48.51	48.28	−2.3	**1**	0	0	**1**
2021	1.63	−46.88	−41.3	0	**1**	0	**1**
2022	1.66	0.03	30.86	0	0	0	0
2023	1.79	0.13	10.49	0	0	0	0
2024	1.88	0.09	25.3	0	0	0	0

Source: Author's Compiled.

Table 26
Descriptive analysis of Tata Sons.

Descriptive Statistics							
	N	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
DE_Ratio	10	5.8980	14.98707	3.151	.687	9.946	1.334
Growth_Assets	10	17.0360	33.09132	.773	.687	3.333	1.334
IndAS	10	.70	.483	−1.035	.687	−1.224	1.334
Pandemic	10	.30	.483	1.035	.687	−1.224	1.334
Suspected_Fraud	10	.20	.422	1.779	.687	1.406	1.334
Emerging_CG	10	.60	.516	−0.484	.687	−2.277	1.334
Valid N (listwise)	10						

Source: Through SPSS compiled by the Author.

Table 27
T-Test of Tata Sons.

One-Sample Test						
	Test Value = 2.262					
	t	df	Sig. (2-tailed)	Mean Difference	95 % Confidence Interval of the Difference	
					Lower	Upper
DE_Ratio	.767	9	.463	3.63600	−7.0851	14.3571
Growth_Assets	1.412	9	.192	14.77400	−8.8981	38.4461
IndAS	−10.226	9	.000	−1.562	−1.91	−1.22
Pandemic	−12.844	9	.000	−1.962	−2.31	−1.62
Suspected_Fraud	−15.465	9	.000	−2.062	−2.36	−1.76
Emerging_CG	−10.178	9	.000	−1.662	−2.03	−1.29

Source: Through SPSS compiled by the Author.

Table 28
Combined multinomial logit model estimation results (2015–2024).

Variable	Model 1: IndAS (β, p-value)	Model 2: Pandemic (β, p-value)	Model 3: Suspected Fraud (SF) (β, p-value)	Model 4: Emergence of Corporate Governance (ECG) (β, p-value)
Debt/Equity Ratio	1.2598, 0.0060	0.1198, 0.2740	–0.0397, 0.2590	1.5122, 0.0098
Growth Assets Rate	–0.0018, 0.9039	–0.0630, 0.1896	–10.6839, 0.0000	–0.0195, 0.2926

Source: Author's Compiled.

Communications (RCom). Before adopting Ind AS, companies recognised impairment losses conservatively, relying on historical cost accounting principles under Indian GAAP, which limited forward-looking evaluations of asset recoverability [39]. In contrast, the introduction of Ind AS 36 required a more thorough and standardised process for impairment testing, mandating companies to determine the recoverable amount of assets at every reporting date and to recognise impairment losses whenever the carrying amount surpasses the recoverable amount.

Table 30 reveals that Reliance Communication experienced significant financial distress from 2018 onward, as indicated by a sharp decline in Return on Equity (RoE) of –856.59 % in 2018 and recurring low Return on Capital Employed (RoCE) values. The implementation of Ind AS began in 2018, aligning with the emergence of corporate governance and the detection of suspected fraud. The pandemic had a further impact on operations from 2020 to 2022. Following 2022, governance efforts continued, but financial recovery remained limited, with stagnant returns.

Table 31 shows evidence of suspected fraud at Reliance Communications for the years 2018, 2019, 2023, and 2024. In 2018, Rule 1 came into effect as the Return on Equity (RoE) plummeted to a negative value of below –100 %. Then, in 2019, Rule 2 was activated due to a decrease in Return on Capital Employed (RoCE) exceeding 20 %. By 2023 and 2024, Rule 3 was triggered when RoE remained at zero for more than three consecutive years. These trends suggest potential financial instability, as well as the possibility of misreporting or governance failures.

4.5.1. *Corporate governance*

Between 2016 and 2024, Reliance Communications experienced fluctuating responses in corporate governance. Significant governance changes emerged in 2018, 2019, 2023, and 2024, coinciding with severe financial distress (e.g., a Return on Equity of –856.59 in 2018), the implementation of IndAS, or allegations of fraud. These years likely prompted either regulatory or internal structural reforms. In contrast, the years 2020 to 2022, despite the effects of the pandemic and ongoing compliance with IndAS, did not see any governance developments, perhaps due to stabilisation efforts or a lack of compelling triggers. Overall, it appears that corporate governance at Reliance

Table 29
Model fit statistics summary of Tata Sons.

Model	Log-Likelihood	AIC	BIC	HQ	Correct Predictions	Likelihood Ratio Test (χ^2, p-value)
Model 1 (IndAS)	–4.29	12.58	13.18	11.91	70 %	5.287 (0.0711)
Model 2 (Pandemic)	–4.24	12.48	13.08	11.81	90 %	5.386 (0.0677)
Model 3 (SF)	–2.44e–08	4.00	4.61	3.34	100 %	13.863 (0.0010)
Model 4 (ECG)	–4.36	12.73	13.33	12.06	70 %	5.135 (0.0767)

Source: Author's Compiled.

Table 30
Comparative analysis impairment losses (Ind AS 36).

Year	RoE	RoCE	IndAS	Pandemic	Suspected Fraud	Emergence of Corporate Governance
2016	2.02	0.84	0	0	0	0
2017	−4.91	0.63	0	0	0	0
2018	−856.59	0.8	1	0	1	1
2019	0	−28.88	1	0	1	1
2020	−88.6	1.13	1	1	0	0
2021	0	0.29	1	1	0	0
2022	0	0.31	1	1	0	0
2023	0	0.17	1	0	1	1
2024	0	0.08	1	0	1	1

Source: Author's calculated (http://www.relianceada.com/reliance-communications).

Table 31
Suspected fraud analysis of reliance communications.

Year	RoE	RoCE	Rule 1 (RoE < −100)	Rule 2 (ΔRoCE < −20)	Rule 3 (3+ yrs RoE = 0)	Fraud_Suspected
2016	2.02	0.84	0	–	0	0
2017	−4.91	0.63	0	Δ = −0.21	0	0
2018	−856.59	0.80	**1**	Δ = +0.17	0	**1**
2019	0	−28.88	0	**Δ = −29.68**	0	**1**
2020	−88.60	1.13	0	**Δ = +30.01**	0	0
2021	0	0.29	0	Δ = −0.84	0	0
2022	0	0.31	0	Δ = +0.02	0	0
2023	0	0.17	0	Δ = −0.14	**1** (3 yrs RoE = 0)	**1**
2024	0	0.08	0	Δ = −0.09	**1** (4 yrs RoE = 0)	**1**

Source: Author's Compiled.

Communications is reactive, primarily influenced by crises, such as fraud and significant losses, rather than being driven by proactive growth or compliance initiatives.

Table 32 indicates possible fraud at Reliance Communications for the years 2018, 2019, 2023, and 2024. In 2018, a significant drop in the Return on Equity (RoE) triggered Rule 1, as it fell below –100 %. In 2019, Rule 2 was activated due to a decline in Return on Capital Employed (RoCE) exceeding 20 %. Subsequently, in 2023 and 2024, Rule 3 was applied because RoE remained at zero for over three consecutive years. These patterns indicate financial instability and raise concerns about potential misreporting or governance issues.

The T-test results in Table 33 show that RoE and RoCE are not statistically significant ($p > 0.05$), indicating no substantial deviation from the test value. However, IndAS, Pandemic, Suspected Fraud, and Emergence of Corporate Governance are all highly significant ($p = 0.000$), with negative t-values, confirming substantial deviations from the mean. This suggests that these factors had a statistically significant impact on Reliance Communication's performance and reporting practices during the observed period, reflecting central governance and operational concerns.

Table 34 presents the multinomial logit results for Reliance Communications, indicating that return on equity (RoE) has a significant impact on all four models. The data indicate a negative impact of RoE on IndAS adoption, suspected fraud, and corporate governance development ($p < 0.05$), suggesting that low equity returns contribute to heightened financial disclosure reforms and governance challenges. Furthermore, RoCE significantly influences suspected fraud and governance ($p = 0.0179$), indicating that diminished capital efficiency escalates concerns and spurs governance responses [40]. These results highlight the impact of financial pressure on driving regulatory and ethical adjustments in 2016.

Table 35 illustrates that Models 3 and 4 (Suspected Fraud and Corporate Governance) provide the best fit, characterised by the lowest AIC, BIC, and HQ values, along with the highest prediction accuracy rates at 77.8 %. Their likelihood ratio tests are statistically significant ($p = 0.0261$), indicating robust explanatory power. In contrast, models 1 and 2 display weaker fits and inferior prediction accuracy, suggesting they are less effective in elucidating the adoption of IndAS and the pandemic's impacts on Reliance Communications.

4.6. Jet airways - related party disclosures (Ind AS 24)

The lack of data for Jet Airways post-2018 largely stems from the company's significant financial turmoil, which resulted in the halt of its operations in April 2019. Jet Airways struggled with an escalating debt burden, rising operational costs, fierce market competition, and decreasing revenues. As a result, the company failed to meet its obligations to lenders, employees, and vendors. These financial challenges

Table 32
Descriptive analysis of reliance communication.

Descriptive Statistics							
	N	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
RoE	9	−105.3422	283.229	−2.942	.717	8.722	1.400
RoCE	9	−2.7367	9.80998	−2.993	.717	8.967	1.400
IndAS	9	.78	.441	−1.620	.717	.735	1.400
Pandemic	9	.33	.500	.857	.717	−1.714	1.400
Suspected_fraud	9	.44	.527	.271	.717	−2.571	1.400
Emergence_Corporate_Governance	9	.44	.527	.271	.717	−2.571	1.400
Valid N (listwise)	9						

Source: Through SPSS compiled by the Author.

Table 33
T-Test analysis of reliance communication.

One-Sample Test						
	Test Value = 2.306					
	t	df	Sig. (2-tailed)	Mean Difference	95 % Confidence Interval of the Difference	
					Lower	Upper
RoE	−1.140	8	.287	−107.64822	−325.3577	110.0613
RoCE	−1.542	8	.162	−5.04267	−12.5833	2.4979
IndAS	−10.397	8	.000	−1.528	−1.87	−1.19
Pandemic	−11.836	8	.000	−1.973	−2.36	−1.59
Suspected fraud	−10.596	8	.000	−1.862	−2.27	−1.46
Emergence Corporate Governance	−10.596	8	.000	−1.862	−2.27	−1.46

Source: Through SPSS compiled by the Author.

Table 34
Combined multinomial logit model estimation results (2016–2024).

Variable	Model 1: IndAS(β, p-value)	Model 2: Pandemic Impact(β, p-value)	Model 3: Suspected Fraud(β, p-value)	Model 4: Emergence of Corporate Governance(β, p-value)
RoE	−−0.0586, 0.0195	0.0036, 0.0201	–0.0122, 0.0038	–0.0122, 0.0038
RoCE	−−0.9898, 0.5145	0.2045, 0.4684	–4.9000, 0.0179	–4.9000, 0.0179

Source: Author's Compiled.

ultimately led Jet Airways to enter insolvency proceedings under the Insolvency and Bankruptcy Code (IBC), 2016.

After 2018, Jet Airways ceased regular publication of its financial statements, attributed to the suspension of operations, ongoing legal struggles, and the insolvency resolution process. Therefore, information regarding its financial performance, such as operating cash flow, debt-equity ratio, and profitability, is unavailable after 2018.

Table 36 illustrates that Jet Airways experienced considerable financial instability from 2009 to 2018, marked by inconsistent growth in Operating Cash Flow (OCF) and negative debt-to-equity ratios. There were indications of fraud in six out of ten years, especially during periods of low profitability and high debt. Although there were early responses to corporate governance issues, the implementation of Ind AS did not occur until 2017. Improvements in governance typically align with fraud detection, suggesting a reactive approach to compliance with financial and ethical standards rather than a proactive one.

Table 37 presents the fraud detection analysis of Jet Airways from 2009 to 2018, utilising qualitative, rule-based indicators from forensic accounting, such as the Beneish M-Score. This analysis reveals troubling patterns, including red flags such as a notable rise in operating profit margins accompanied by negative or erratic cash flows, as well as highly atypical debt-to-equity ratios in certain years. Specifically, the years 2009, 2010, 2011, 2013, 2016, and 2017 showed combinations of these red flags, suggesting possible earnings manipulation. These findings underscore discrepancies between reported profitability and financial health. While these patterns do not provide conclusive evidence of fraud, they warrant a comprehensive forensic investigation to determine possible misreporting.

Table 36
Comparative analysis of jet airways with related party disclosures.

Year	Growth of OCF	D/E Ratio	OPM (%)	Suspected Fraud	Ind AS	Enhanced Corporate Governance
2009	−194.7	49.69	−3.51	1	0	1
2010	488.81	−167.2	12.25	1	0	1
2011	3.66	−68.33	12.77	1	0	1
2012	32.62	−6.37	2.01	0	0	0
2013	−18.01	−3.65	6.62	1	0	1
2014	−46.22	−1.94	−7.61	0	0	0
2015	−26.71	−1.54	1.61	0	0	0
2016	238.87	−1.67	13.21	1	0	1
2017	−59.5	−1.11	13.11	1	1	1
2018	68.14	−0.74	3.04	0	1	0

Source: Author's calculations with company annual reports (https://www.jetairways.com).

Table 37
Suspected fraud detection.

Year	Growth of OCF	D/E Ratio	OPM (%)	Suspected Fraud (Y/N)?	Remarks
2009	−194.7	49.69	−3.51	**Yes**	Very high D/E, poor margins, significant OCF drop
2010	488.81	−167.2	12.25	**Yes**	Extreme D/E swing, sudden huge OCF spike
2011	3.66	−68.33	12.77	**Yes**	High profit margin with flat OCF
2012	32.62	−6.37	2.01	No	Relatively normal
2013	−18.01	−3.65	6.62	**Yes**	Positive margin but negative OCF.
2014	−46.22	−1.94	−7.61	No	Loss and negative OCF, expected
2015	−26.71	−1.54	1.61	No	Low all around
2016	238.87	−1.67	13.21	**Yes**	High OCF growth and margin, but flat D/E
2017	−59.5	−1.11	13.11	**Yes**	Huge margin, declining cash
2018	68.14	−0.74	3.04	No	Moderate values

Source: Author's Compiled.

Table 35
Model fit statistics summary.

Model	Log-Likelihood	AIC	BIC	HQ	Correct Predictions	Likelihood Ratio Test (χ^2, p-value)
Model 1: IndAS	–4.1027	12.2054	12.5999	11.3542	55.6 % (5/9)	4.271, 0.1182
Model 2: Pandemic	–5.0682	14.1363	14.5308	13.2851	44.4 % (4/9)	2.340, 0.3103
Model 3: Suspected Fraud	–2.5926	9.1852	9.5796	8.3340	77.8 % (7/9)	7.291, 0.0261
Model 4: Corporate Governance	–2.5926	9.1852	9.5796	8.3340	77.8 % (7/9)	7.291, 0.0261

Source: Author's Compiled.

Table 38
Descriptive analysis of jet airways.

Descriptive Statistics							
	N	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Growth OCF	10	48.6960	189.33334	1.530	.687	2.898	1.334
Debt / Equity	10	−20.2860	58.73037	−1.987	.687	4.758	1.334
Operating PM	10	5.3500	7.47750	−0.426	.687	−1.026	1.334
Suspected Fraud	10	.60	.516	−0.484	.687	−2.277	1.334
IndAS	10	.20	.422	1.779	.687	1.406	1.334
Enhanced Corporate Governance	10	.60	.516	−0.484	.687	−2.277	1.334
Valid N (list wise)	10						

Source: Through SPSS compiled by the Author.

4.6.1. Enhanced corporate governance

Jet Airways needed to implement stronger corporate governance from 2009 to 2017 to avert financial strain and enhance accountability. During these years, indications of earnings manipulation or misreporting risks were present, which better board oversight, robust internal audits, and compliance with corporate governance codes (like Clause 49 or SEBI LODR) could have addressed.

Table 38 presents a descriptive analysis of Jet Airways from 2009 to 2018, revealing considerable volatility in Operating Cash Flow (OCF), marked by a significant standard deviation of 189.33 and a positive skewness of 1.53, which indicates occasional sharp spikes. The average Debt-to-Equity ratio is significantly negative at −20.29, exhibiting high variability. Its skewness of −1.99 and kurtosis of 4.76 indicate extreme debt levels during certain years. The Operating Profit Margin (OPM) remains modest at a mean of 5.35 %. Fraud was suspected in 60 % of the years examined. The adoption of Ind AS was low at 20 %, whereas corporate governance practices improved in 60 % of the observed period, suggesting that these improvements were primarily motivated by financial or compliance pressures.

Table 39 presents the *t*-test results, demonstrating that there are no significant differences in Growth_OCF, Debt-Equity, and Operating Profit Margin about the test value, as their p-values exceed 0.05. On the other hand, the factors of Suspected Fraud, Ind AS adoption, and Corporate Governance measures show statistical significance ($p = 0.000$), indicating that these aspects substantially underperformed against the benchmark. This highlights Jet Airways' inadequate compliance and governance practices, which may stem from reactive responses to financial challenges or irregularities.

The multinomial logit model for Jet Airways (2009–2018), presented in Table 40, reveals limited statistical significance among the variables, indicating weak connections between financial metrics and governance actions. In Model 3, variables including Growth of Operating Cash Flow (OCF), Debt-Equity Ratio (DER), and Operating Profit Margin (OPM) display high p-values (above 0.36), showing a negligible effect on Ind AS adoption. Likewise, in Models 4 and 5, while OPM has a moderately strong positive coefficient ($\beta = 0.3599$), its p-value of 0.1101 shows that it lacks statistical significance. This implies that profitability and financial structure had a minimal influence on fraud detection and the effectiveness of enhanced corporate governance measures. Overall, Jet Airways appears to have established governance and compliance mechanisms primarily in response to crises rather than proactively, based on its financial health and operational efficiency.

Table 40
Combined multinomial logit model estimation results (2009–2018).

Variable	Model 3: IndAS (β, p-value)	Model 4: Suspected Fraud (β, p-value)	Model 5: ECG (β, p-value)
Growth of OCF	0.0010, 0.8758	−−0.0135, 0.3575	−−0.0135, 0.3575
Debt-Equity Ratio (DER)	0.0170, 0.3601	–0.0230, 0.4424	–0.0230, 0.4424
Operating Profit Margin	–0.0134, 0.8896	0.3599, 0.1101	0.3599, 0.1101

Source: Author's Compiled.

Table 41 illustrates that Models 4 and 5 (Suspected Fraud and Enhanced Corporate Governance) provide a better fit, as indicated by lower AIC, BIC, and HQ values, coupled with a higher correct prediction rate of 70 %. However, none of the models achieve statistical significance ($p > 0.05$), suggesting a limited ability to explain the governance or fraud outcomes at Jet Airways.

4.7. Vodafone idea – contingent liabilities (Ind AS 37)

From 2016 to 2024, Vodafone Idea experienced a notable increase in contingent liabilities, primarily due to regulatory dues and legal disputes, particularly following the Supreme Court's AGR ruling. The company adopted a transparent disclosure of these liabilities under Ind AS 37, recognising potential outflows without classifying them as provisions. This method enhanced clarity regarding financial risks while highlighting Vodafone Idea's ongoing legal and economic challenges,

Table 39
T-Test of Jet Airlines.

One-Sample Test						
	Test Value = 2.262					
	t	df	Sig. (2-tailed)	Mean Difference	95 % Confidence Interval of the Difference	
					Lower	Upper
Growth OCF	.776	9	.458	46.43400	−89.0069	181.8749
Debt/Equity	−1.214	9	.256	−22.54800	−64.5612	19.4652
Operating PM	1.306	9	.224	3.08800	−2.2611	8.4371
Suspected Fraud	−10.178	9	.000	−1.662	−2.03	−1.29
IndAS	−15.465	9	.000	−2.062	−2.36	−1.76
Enhanced Corporate Governance	−10.178	9	.000	−1.662	−2.03	−1.29

Source: Through SPSS compiled by the Author.

Table 41
Model fit statistics summary.

Model	Log-Likelihood	AIC	BIC	HQ	Correct Predictions	Likelihood Ratio Test (χ^2, p-value)
Model 3: IndAS	–6.254	18.509	19.416	17.513	6/10 (60 %)	1.354, 0.716
Model 4: Suspected Fraud	–4.161	14.322	15.230	13.326	7/10 (70 %)	5.541, 0.136
Model 5: ECG	–4.161	14.322	15.230	13.326	7/10 (70 %)	5.541, 0.136

Source: Author's Compiled.

Table 42
Descriptive analysis of Vodafone Idea.

Year	Growth % % Contingent Liability	Debt/Equity Ratio	Interest Coverage Ratio	Suspected Fraud (1) / Not (0)	Pandemic	IndAS	Emergence of Corporate Governance
2016	0	1.59	3.33	0	0	0	0
2017	100	2.09	0.68	1	0	1	1
2018	0	2.09	−0.42	1	0	1	1
2019	31.93	1.82	−1.03	1	0	1	1
2020	−5.14	16.11	0.67	1	1	1	1
2021	37.16	−4.12	0.73	1	1	1	1
2022	−8.36	−3.08	0.77	1	1	1	1
2023	18.2	−18	−0.36	1	0	1	1
2004	−15.41	−1.99	−0.55	1	0	1	1

Source: Author's calculations with company annual reports (https://www.myvi.in).

particularly in addressing its substantial government dues [41].

Table 42's analysis of Vodafone Idea from 2016 to 2023 shows significant financial distress and governance changes. Rising contingent liabilities and fluctuating debt-to-equity ratios indicate instability, while low or negative interest coverage ratios reveal challenges in meeting debt obligations. Since 2017, suspicions of fraud have coincided with the implementation of Ind AS and improvements in governance, reflecting a reactive strategy. Financial pressure escalated during the pandemic (2020–2022), despite the ongoing implementation of governance frameworks. While there were periods of growth, the overall trend highlights financial vulnerability and compliance driven by necessity.

To evaluate possible fraud using financial indicators, the Fraud Indicators Heuristic (Rule-based logic), as shown in Table 43, is utilised. Important metrics include growth in contingent liabilities, the Debt/Equity Ratio, and the Interest Coverage Ratio, all as per Ind AS 37. A year is marked as potentially fraudulent (1) if any of the subsequent conditions are met: contingent liabilities increase by over 30 %, the Debt/Equity Ratio is negative or excessively skewed, or the Interest Coverage Ratio falls below a certain threshold.

4.7.1. Corporate governance

Vodafone Idea's corporate governance practices appear to be reactive, arising from financial alerts and regulatory pressures. Since 2017, the company has demonstrated consistent compliance (rated as 1), likely due to the increasing presence of contingent liabilities, significant debt, and persistent indications of potential fraud. This shift suggests a strategic move towards governance reforms aimed at addressing stakeholder concerns and regulatory scrutiny, particularly following the adoption of Ind AS. In contrast, the year 2016 shows no indication of governance efforts (rated as 0) despite a relatively stable financial health. Thus, it seems that governance at Vodafone Idea is driven more by necessity than by a proactive commitment to corporate responsibility.

Table 43
Suspected fraud analysis.

Year	Growth % CL	D/E Ratio	ICR	Fraud Conditions Met?	Suspected Fraud (1/0)
2016	0	1.59	3.33	None	0
2017	100	2.09	0.68	High CL growth	1
2018	0	2.09	−0.42	Negative ICR	1
2019	31.93	1.82	−1.03	High CL, Negative ICR	1
2020	−5.14	16.11	0.67	Abnormal D/E	1
2021	37.16	−4.12	0.73	High CL, Negative D/E	1
2022	−8.36	−3.08	0.77	Negative D/E	1
2023	18.2	−18	−0.36	Negative D/E, ICR	1
2004	−15.41	−1.99	−0.55	Negative D/E, ICR	1

Source: Author's Compiled.

The analysis of Vodafone Idea in Table 44 reveals significant variability in contingent liability growth (mean = 17.60, SD = 35.84) and interest coverage (mean = 0.42), suggesting financial instability. The data shows a positive skew, indicating that there are frequent extreme values. Factors such as suspected fraud, Ind AS, and corporate governance exhibit high negative skewness and peaked kurtosis, indicating consistent reporting over the years. The ongoing indications of fraud and governance reforms underscore a reactive approach to compliance in the face of ongoing financial pressures.

Table 45 presents the *t*-test results for Vodafone Idea, showing that the interest coverage ratio ($p = 0.002$), indications of fraud, the pandemic's impact, Ind AS adoption, and the enhancement of corporate governance (all with $p = 0.000$) are statistically significant. This highlights noteworthy deviations from the expected value. The findings suggest that Vodafone Idea encountered financial difficulties, regulatory hurdles, and ongoing signs of fraud. In contrast, the increase in contingent liabilities and the debt-equity ratio did not exhibit statistical significance, indicating that these elements did not reliably influence the results.

The multinomial logit model (2015–2024) for Vodafone Idea, outlined in Table 46, reveals that none of the predictors across all assessed models are statistically significant (p-values > 0.05). The increase in contingent liabilities ($\beta = 0.0485$, $p = 0.1161$) demonstrates a moderate but insignificant correlation with suspected fraud, Ind AS, and ECG. Similarly, the debt-equity and interest coverage ratios exhibit weak effects. In conclusion, financial metrics had a negligible impact on changes in governance, fraud detection, or responses to the pandemic, suggesting that reforms were primarily driven by external pressures rather than internal financial factors

In Table 47, the financial predictors for Models 2, 5, 8, and 13—Growth in Contingent Liabilities, Debt/Equity Ratio, and Interest Coverage Ratio—are consistently statistically insignificant ($p > 0.05$).

Table 44
Descriptive analysis of Vodafone Idea.

Descriptive Statistics							
	N	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Growth CL	9	17.5978	35.84388	1.723	.717	3.282	1.400
Debt / Equity	9	−0.3878	8.85421	−0.231	.717	2.866	1.400
Interest Coverage	9	.4244	1.28326	1.486	.717	3.090	1.400
Suspected Fraud	9	.89	.333	−3.000	.717	9.000	1.400
Pandemic	9	.33	.500	.857	.717	−1.714	1.400
IndAS	9	.89	.333	−3.000	.717	9.000	1.400
Emergence CG	9	.89	.333	−3.000	.717	9.000	1.400
Valid N (list wise)	9						

Source: Through SPSS compiled by the Author.

Table 45
T-Test of Vodafone idea.

One-Sample Test						
	Test Value = 2.306					
	t	df	Sig. (2-tailed)	Mean Difference	95 % Confidence Interval of the Difference	
					Lower	Upper
Growth CL	1.280	8	.236	15.29178	−12.2603	42.8438
Debt/Equity	−0.913	8	.388	−2.69378	−9.4997	4.1122
Interest Coverage	−4.399	8	.002	−1.88156	−2.8680	−0.8952
Suspected Fraud	−12.754	8	.000	−1.417	−1.67	−1.16
Pandemic	−11.836	8	.000	−1.973	−2.36	−1.59
IndAS	−12.754	8	.000	−1.417	−1.67	−1.16
Emergence CG	−12.754	8	.000	−1.417	−1.67	−1.16

Source: Through SPSS compiled by the Author.

Table 46
Combined multinomial logit model estimation results (2015–2024).

Variable	Model 13: ECG ⟨br⟩ (β, p-value)	Model 8: IndAS ⟨br⟩ (β, p-value)	Model 2: Suspected Fraud ⟨br⟩ (β, p-value)	Model 5: Pandemic Impact ⟨br⟩ (β, p-value)
Growth in Contingent Liability	0.0485, 0.1161	0.0485, 0.1161	0.0485, 0.1161	–0.0192, 0.4041
Debt/Equity Ratio	0.0215, 0.8115	0.0215, 0.8115	0.0215, 0.8115	0.0702, 0.4920
Interest Coverage Ratio	–0.6240, 0.2482	–0.6240, 0.2482	–0.6240, 0.2482	0.0565, 0.9171

Source: Author's Compiled.

While a weak positive correlation exists between growth in contingent liabilities and suspected fraud and compliance, it remains statistically insignificant across all models. Each model, except the one related to the pandemic (44.4 %), shows moderate overall predictive accuracy (66.7 %), highlighting the pandemic's impact as an external shock. In conclusion, the results suggest that these financial ratios do not significantly account for IndAS compliance, governance emergence, or fraud detection, emphasising the likely greater impact of regulatory mandates compared to economic aspects.

4.8. SAIL (Steel Authority of India) - Asset depreciation (Ind AS 16)

The introduction of Ind AS 16 — Property, Plant, and Equipment — represented a significant change in the depreciation policy and accounting practices for assets among Indian companies, including Steel Authority of India Limited (SAIL). Before this standard, depreciation was primarily determined through fixed rates and prescribed schedules, in accordance with the Indian Generally Accepted Accounting Principles (GAAP). In contrast, Ind AS 16 brought a more flexible method, focusing on the componentisation of assets, aligning depreciation practices with the patterns of economic benefits, and requiring regular assessments of the useful life and residual value of assets.

Table 48 provides a comparative analysis of SAIL from 2016 to 2024, highlighting erratic net income growth characterised by significant declines in 2016, 2018, and 2023, often coinciding with suspected years of fraud. Despite a consistent application of Ind AS after 2017, improvements in governance (ECG) showed variability. The notably high depreciation and changing asset turnover indicate potential financial instability. Reports of fraud appear to align with discrepancies in asset efficiency and profit reporting, suggesting possible manipulation or mismanagement.

Table 49 shows a pattern of suspected fraud over several years, highlighting significant financial discrepancies. Suspicions arose in 2016, 2018, 2020, and 2023, coinciding with considerable warning signs like drastic net income drops (−292 % in 2016, −141.07 % in 2018, and −82.21 % in 2023) and unusually high depreciation and amortisation increases (40.15 % in 2020). The 2023 spike in the Fixed Asset Turnover Ratio to 62 suggests possible manipulation or reporting inconsistencies. These indicators support the need for forensic investigations and governance scrutiny to ensure accountability.

4.8.1. Corporate governance

The Emergence of Corporate Governance (ECG) happens in years when key governance indicators align, specifically, the lack of suspected fraud, the implementation of Ind AS, and improved financial metrics. ECG was evident in 2017, 2019, 2021, 2022, and 2024, which can be

Table 47
Model fit statistics summary.

Model	Log-Likelihood	AIC	BIC	HQ	Correct Predictions	Likelihood Ratio Test (χ^2, p-value)
Model 13: ECG	–4.5196	15.0393	15.6310	13.7625	6 (66.7 %)	3.437 (0.3290)
Model 8: IndAS	–4.5196	15.0393	15.6310	13.7625	6 (66.7 %)	3.437 (0.3290)
Model 2: SF	–4.5196	15.0393	15.6310	13.7625	6 (66.7 %)	3.437 (0.3290)
Model 5: Pandemic	–5.3997	16.7994	17.3911	15.5226	4 (44.4 %)	1.677 (0.6420)

Source: Author's Compiled.

Table 48
Comparative analysis of SAIL (Steel Authority of India) - Asset depreciation (Ind AS 16).

Year	Fixed Asset Turnover Ratio	D&A Expense Growth %	Net Income Growth %	Suspected Fraud	Pandemic	Ind AS	ECG
2016	0.59	10.1	−292	1	0	0	0
2017	0.64	−8.38	−51.75	0	0	1	1
2018	0.77	24.27	−141.07	1	0	1	0
2019	0.87	7.68	239.31	0	0	1	1
2020	0.82	40.15	−25.32	1	1	1	0
2021	1.33	9.23	95.61	0	1	1	1
2022	0.86	4.2	195.22	0	1	1	1
2023	62	16.1	−82.21	1	0	1	0
2024	0.79	6.32	40.89	0	0	1	1

Source: Author's calculations with company annual reports(https://sail.co.in/en).

Table 49
Suspected Fraud Analysis.

Year	FATR	FATR Justification	D&A Growth %	D&A Justification	Net Income Growth %	NI Justification	Suspected Fraud
2016	0.59	Normal	10.1	Acceptable (< 30 %)	−292	Large drop (> −200 %)	1
2017	0.64	Slight increase	−8.38	Moderate decrease	−51.75	Normal variation	0
2018	0.77	Slight increase	24.27	Acceptable	−141.07	Still a large drop	1
2019	0.87	Moderate increase	7.68	Acceptable	239.31	Recovery noted	0
2020	0.82	Stable	40.15	Sudden high jump >30 %	−25.32	Acceptable dip	1
2021	1.33	High but under the threshold	9.23	Acceptable	95.61	Good performance	0
2022	0.86	Drop but stable	4.2	Acceptable	195.22	High, but not anomalous	0
2023	62	Huge spike (> 50x)	16.1	Acceptable	−82.21	Large reversal	1
2024	0.79	Normal	6.32	Acceptable	40.89	Acceptable	0

Source: Author's Compiled.

Table 50
Descriptive Statistics of SAIL (Steel Authority of India).

Descriptive Statistics							
	N	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
FA_TOR	9	7.6300	20.38983	2.999	.717	8.997	1.400
DA_EG	9	12.1856	13.68333	.888	.717	1.677	1.400
NIG	9	−2.3689	166.35252	−0.166	.717	−0.204	1.400
Suspected Fraud	9	.44	.527	.271	.717	−2.571	1.400
Pandemic	9	.33	.500	.857	.717	−1.714	1.400
IndAS	9	.89	.333	−3.000	.717	9.000	1.400
ECG	9	.56	.527	−0.271	.717	−2.571	1.400
Valid N (listwise)	9						

Source: Through SPSS compiled by the Author.

attributed to better regulatory compliance, stable or increasing profitability, and fewer discrepancies. These years reflect an environment where governance aspects such as transparency, accounting standards, and ethical oversight likely encouraged organisational discipline. Conversely, ECG was absent in years characterised by high volatility, suspected fraud, or substantial income declines, highlighting that ECG is more likely to manifest when internal controls and financial integrity are effectively upheld.

Table 50 provides the descriptive statistics for SAIL, highlighting significant variability in financial metrics. The Fixed Asset Turnover Ratio, averaging 7.63, exhibits considerable skewness (2.999) and kurtosis (8.997), suggesting the presence of outliers and a non-normal distribution. Net Income Growth, with an average of –2.37, shows notable volatility, as evidenced by its high standard deviation. While the adoption of Ind AS is prevalent, averaging 0.89, cases of suspected fraud and ECG scores indicate moderate rates and an inconsistent impact of governance over the observed period.

Table 51 displays the T-test findings for SAIL, revealing no notable differences in the Fixed Asset Turnover Ratio, Depreciation & Amortisation growth, or Net Income Growth compared to the test value. Conversely, factors such as Suspected Fraud, Pandemic, Ind AS adoption, and Enhanced Corporate Governance (ECG) show statistical

Table 51
T-Test of SAIL (Steel Authority of India).

One-Sample Test						
	Test Value = 2.306					
	t	df	Sig. (2-tailed)	Mean Difference	95 % Confidence Interval of the Difference	
					Lower	Upper
FA_TOR	.783	8	.456	5.32400	−10.3490	20.9970
DA_EG	2.166	8	.062	9.87956	−0.6384	20.3975
NIG	−0.084	8	.935	−4.67489	−132.5448	123.1950
Suspected Fraud	−10.596	8	.000	−1.862	−2.27	−1.46
Pandemic	−11.836	8	.000	−1.973	−2.36	−1.59
IndAS	−12.754	8	.000	−1.417	−1.67	−1.16
ECG	−9.964	8	.000	−1.750	−2.16	−1.35

Source: Through SPSS compiled by the Author.

significance ($p < 0.001$), reflecting significant deviations and implying that regulatory and external factors have a substantial impact on financial disclosures and governance practices.

Table 52 of the multinomial logit results indicates that financial

Table 52
Combined multinomial logit model estimation results (2016–2024).

Variable	Model 1: ECG (β, p-value)	Model 2: IndAS (β, p-value)	Model 3: SF (β, p-value)	Model 4: Pandemic Impact (β, p-value)
Fixed Asset Turnover	10.3764, 0.0000	11.4622, 0.0000	−5.78215, 0.0000	−1.54315, 0.2363
D&A Expense Growth	−22.6461, 0.0000	0.0568, 0.2436	12.6803, 0.0000	0.06902, 0.1312
Net Income Growth	3.5726, 0.0000	0.0408, 0.0000	−1.99862, 0.0000	0.00833, 0.2200

Source: Author's Compiled.

variables have a significant impact on governance and reporting. A positive correlation exists between Fixed Asset Turnover and the adoption of ECG and IndAS, although it may impede the identification of suspected fraud. In contrast, an increase in depreciation hurts ECG, raising concerns about potential fraud. Net Income Growth supports both ECG and IndAS while concurrently reducing fraud levels. The pandemic's effect turned out to be statistically insignificant. In conclusion, asset efficiency and profitability strengthen governance, whereas inconsistent depreciation could signal potential fraud risks.

Table 53 displays the model fit statistics, indicating exceptional predictive accuracy (100 %) for ECG, IndAS, and Suspected Fraud, supported by low AIC/BIC values and notable likelihood ratio tests ($p < 0.01$). Conversely, the Pandemic Impact model demonstrates poorer performance, with an accuracy of 77.8 % and an insignificant p-value (0.2709), highlighting its limited explanatory capability.

5. Analysing and interpreting data: validation of data by objectives and empirical evidence

This section confirms the research objectives by providing empirical evidence derived from different Indian corporate case studies. Each objective is evaluated using distinct data points and their analysis, highlighting the significant influence of Indian Accounting Standards (Ind AS) on financial reporting, corporate governance, and investor protection.

Objective 1: Assess how Ind AS, with its potential to enhance the accuracy, transparency, and reliability of financial reports, can significantly reduce financial misreporting and foster a more transparent and reliable financial landscape.

This goal is robustly backed by empirical data from case studies involving IL&FS, Wipro, Yes Bank, and Reliance Communications. The introduction of specific Ind AS standards required companies to depict a more accurate and transparent financial status, effectively tackling previous cases of misreporting. The validation with data is presented in Table 54.

These cases illustrate that although Ind AS 1 at IL&FS encountered issues with intentional manipulation, other standards effectively improved transparency. For instance, Ind AS 115 at Wipro provided more accurate revenue reporting, Ind AS 109 at Yes Bank revealed significant asset quality problems, and Ind AS 36 at Reliance Communications enforced precise asset valuation. Together, they contributed to a more transparent and trustworthy financial reporting landscape.

Objective 2: Explore how Ind AS adoption can revolutionise corporate governance by enforcing stricter disclosure norms and accountability, and enhancing transparency and ethical behaviour.

The implementation of Ind AS has a significant impact on corporate

Table 55
Validation with data (Objective 2).

Company	Applied Ind AS	Corporate Governance Impact
Jet Airways	Ind AS 24	Failing to disclose related party transactions significantly contributed to the financial collapse, underscoring the vital importance of strict compliance in governance.
Tata Sons	Ind AS 110	Mandatory consolidation eliminated off-balance-sheet liabilities, enhancing financial transparency for the group and bolstering governance oversight.
Vodafone Idea	Ind AS 37	The disclosure of contingent liabilities related to AGR dues, totalling ₹41,202 crore in 2019, increased transparency, informed stakeholders about potential risks, and fostered accountability.

Source: Author's Compiled.

Table 53
Model fit statistics summary.

Model	Log-Likelihood	AIC	BIC	HQ	Correct Predictions	Likelihood Ratio Test (χ^2, p-value)
ECG (Model 1)	−0.000024	6.0000	6.5917	4.7232	9 (100 %)	12.477 (0.0059)
Ind AS (Model 2)	−0.031	6.06	6.65	4.78	100 %	12.416 (0.0061)
Suspected Fraud (Model 3)	−0.0036	6.01	6.60	4.73	100 %	12.469 (0.0059)
Pandemic Impact (Model 4)	−4.281	14.56	15.15	13.29	77.8 %	3.914 (0.2709)

Source: Author's Compiled.

Table 54
Validation with data (Objective 1).

Company	Applied Ind AS	Financial Misreporting Identified	Impact of Ind AS
IL&FS	Ind AS 1	Management of liabilities, concealed debt, and exaggerated Current Ratio (rising from 7.4 % in 2017 to 318.2 % in 2024).	Mandatory reporting of actual financial status, despite ongoing irregularities caused by intentional manipulation.
Wipro	Ind AS 115	Revenue manipulation is regulated.	Revenue growth became more achievable and aligned with actual customer contracts, ensuring precise earnings recognition and accuracy.
Yes Bank	Ind AS 109	Exaggerated asset quality before Ind AS.	The recognition of NPAs surged significantly from 1.28 % in 2018 to 16.8 % in 2020, indicating the actual condition of the assets.
Reliance Communications	Ind AS 36	Impairment losses had not been reported previously.	ROE decreased from 2.02 % in 2016 to −856.59 % in 2018 following the recognition of impairment, providing a true reflection of the economic situation.

Source: Author's Compiled.

governance by mandating increased disclosure and accountability, which fosters more ethical corporate conduct. This transformation is illustrated by various case studies, which show that specific Ind AS standards either necessitated enhanced transparency or revealed governance shortcomings. The validation with data is presented in Table 55.

The cases demonstrate that adhering to Ind AS 24 could have prevented governance issues at Jet Airways. Tata Sons' compliance with Ind AS 110 showcases proactive governance via consolidation. Vodafone Idea's Ind AS 37 disclosures highlighted substantial liabilities and enhanced accountability by explicitly defining emerging risks. These instances underscore the vital importance of Ind AS in promoting better disclosure and accountability, which ultimately leads to improved corporate governance.

Objective 3: Conduct a thorough analysis of the practical challenges businesses may face during the implementation of Ind AS, including system overhauls and employee training.

Implementing Ind AS presents significant practical challenges for businesses owing to its complexity and adherence to international standards. Frequent problems include difficulties in data handling, system integration, and ensuring adequate employee expertise, as demonstrated by the experiences of various companies.

5.1. Challenges identified across case studies

a) Vodafone Idea experienced significant fluctuations in its contingent liability disclosures, which increased from zero in 2017 to ₹41,202 crore in 2019, per Ind AS 37. This suggests difficulties in consistently identifying, measuring, and reporting these complex items, indicating a need for substantial system upgrades and specialised training.
b) Jet Airways encountered notable difficulties in reporting lease obligations and related-party disclosures following the adoption of Indian Accounting Standards (Ind AS). This highlights the challenges of modifying existing systems and training staff to comply with the rigorous demands of new standards, such as Ind AS 116 (Leases) and Ind AS 24 (Related Party Disclosures).
c) Following the implementation of Ind AS 16, SAIL saw significant fluctuations in depreciation expenses, with a 40.15 % rise in the depreciation growth rate in 2020. This indicates difficulties in reassessing asset lifespans, adopting different depreciation methods, and achieving precise system calculations, all of which require considerable system enhancements and re-training of accounting staff.

These examples illustrate that effectively adopting Ind AS demands significant investment in IT infrastructure, restructuring financial processes, and comprehensive training initiatives. Such actions are crucial to bridging the knowledge gap and ensuring accurate data capture and reporting within the new framework.

Objective 4: Investigate how the adoption of Ind AS enhances investor protection by promoting transparency and ethical corporate behaviour.

The implementation of Ind AS significantly enhances investor safeguards by promoting increased transparency and encouraging ethical conduct among corporations, enabling investors to make more informed choices. This goal is supported by results seen in numerous companies, where disclosures required by Ind AS offered vital understanding of financial status and risks. The validation with data is presented in Table 56.

These examples collectively demonstrate that consistent and reliable reporting (Wipro), vital risk disclosures (Yes Bank, Vodafone Idea), or comprehensive transparency at the group level (Tata Sons) allow Ind AS to provide investors with timely and accurate information essential for protecting their interests and assessing corporate integrity.

Table 56
Validation with data (Objective 4).

Company	Investor Protection Outcome Post Ind AS
Wipro	Transparent revenue recognition methods boosted profitability by providing investors with reliable earnings data and reducing information asymmetry.
Yes Bank	The introduction of Ind AS 109 significantly emphasised NPA and provisioning data, improving investor awareness and risk perception, and allowing for accurate risk pricing.
Tata Sons	Under Ind AS 110, consolidated financial reporting increases reliability for investors by providing a comprehensive picture of group-level assets and liabilities, thereby minimising concealed risks.
Vodafone Idea	Disclosure of AGR-related contingent liabilities (under Ind AS 37) informed stakeholders of potential significant risks, enabling a more accurate assessment of the company's financial vulnerability.

Source: Author's Compiled.

Objective 5: Evaluate case studies of companies adopting Ind AS to reduce financial shenanigans and fraudulent reporting.

The analysis of various case studies shows that adopting Ind AS greatly helps diminish financial misconduct and fraudulent reporting. It achieves this by discouraging these activities and facilitating their detection. Relevant data can be found in Table 57.

Results from various companies highlight the significant impact of Ind AS. It has not only prevented revenue manipulation at Wipro and promoted accurate asset valuations at Reliance Communications and SAIL, but also uncovered hidden liabilities at Yes Bank and enabled transparent reporting at the group level for Tata Sons. Ind AS has emerged as a strong deterrent against various forms of financial misconduct. In situations like IL&FS and Jet Airways, where major issues arose, the principles of Ind AS provided a framework for detecting misreporting, thereby fostering greater accountability. This extensive empirical evidence reinforces the crucial role of Ind AS in promoting corporate accountability and reducing the likelihood of financial wrongdoing in India's corporate sector.

6. Hypothesis-wise validation

This study's validation of each hypothesis offers strong empirical insights into how Indian Accounting Standards (Ind AS) affect corporate

Table 57
Validation with data (Objective 5).

Companies Covered	Evidence of Fraud Prevention/Exposure Post-Ind AS Implementation
IL&FS, Jet Airways, Wipro, Yes, Bank, Tata Sons, Vodafone Idea, Reliance Communications, SAIL	The introduction of Ind AS resulted in stricter disclosure requirements, including those concerning related parties at Jet Airways and contingent liabilities at Vodafone Idea. It required precise impairment reporting for companies like RCom and SAIL, ensured accurate revenue recognition at Wipro, necessitated the consolidation of subsidiaries such as Tata Sons, and uncovered concealed financial instruments and bad debts at Yes Bank. The trends in ratios observed post-Ind AS implementation, such as realistic non-performing assets (NPAs), actual asset values, and stable revenues, consistently reflected enhanced financial discipline and integrity, making it significantly harder to hide fraudulent reporting.

Source: Author's Compiled.

financial reporting, governance, and operational efficiency.

Hypothesis 1 (H1)*: The adoption of Ind AS significantly reduces financial shenanigans by enhancing financial reporting accuracy, transparency, and accountability.*

This hypothesis is widely accepted due to robust evidence from various case studies. For example, in the case of IL&FS, liabilities were misclassified, thereby inflating current ratios prior to the implementation of Ind AS. The roll-out of Ind AS 1 introduced stricter classification standards, which effectively reduced such misreporting. In a similar vein, Yes Bank experienced a significant increase in gross Non-Performing Assets (NPAs) following the enforcement of Ind AS 109, which required accurate asset classification, thereby revealing previously hidden bad loans. Additionally, Reliance Communications faced significant asset impairment disclosures under Ind AS 36, resulting in a marked decrease in Return on Equity (ROE), which accurately reflected the company's actual financial status. These examples strongly support H1 by illustrating how Ind AS has played a crucial role in uncovering and mitigating financial misstatements.

Hypothesis 0 (H0): *There is no significant relationship between the adoption of Ind AS and the reduction of financial malfeasance in financial reporting.*

This hypothesis was eventually dismissed. Evidence collected from different firms contradicts this null hypothesis, as each case consistently highlights the role of Ind AS in revealing discrepancies and improving reporting standards. The data-driven dismissal of H0 affirms that Ind AS significantly contributes to fostering financial integrity in corporate India.

Hypothesis 2 (H2): *Ind AS improves corporate governance and investor protection by enforcing stricter disclosure norms and ethical behaviour.*

Compelling examples further support this hypothesis. The downfall of Jet Airways was partly linked to the failure to disclose related-party transactions, a problem that adherence to Ind AS 24 could have helped avoid. Conversely, Tata Sons benefited from Ind AS 110, which mandated the consolidation of subsidiaries, thus preventing the hiding of liabilities and enhancing governance practices. In a similar vein, Vodafone Idea's compliance with Ind AS 37 ensured complete transparency regarding substantial contingent liabilities, protecting investors from unexpected financial shocks. These cases illustrate that Ind AS not only enhances accounting accuracy but also promotes corporate ethics and strengthens investor confidence.

Hypothesis 3 (H3): *Challenges in adopting Ind AS, such as system changes and employee training, negatively impact its implementation effectiveness.*

The difficulties in implementing Ind AS, especially concerning system upgrades and employee training, were also noted. While the long-term advantages of Ind AS are clear, its rollout has faced challenges. For example, Vodafone Idea encountered inconsistent reporting of contingent liabilities and a heavier compliance load during the initial adoption phase. Jet Airways struggled with the implementation of new lease accounting and related party disclosure rules, resulting in operational strain during the transition. Similarly, SAIL saw a considerable rise in depreciation expenses after implementing Ind AS 16, underscoring the financial impact of accurate asset valuation. These instances demonstrate that while Ind AS enhances transparency, it also presents significant transitional hurdles that can influence short-term performance and compliance demands.

Hypothesis 4 (H4): *Companies that successfully implement Ind AS experience fewer financial anomalies than those that do not.*

This hypothesis was confirmed as well. Wipro exemplifies this by maintaining steady profit margins through accurate revenue recognition under Ind AS 115, effectively reducing financial discrepancies. Similarly, Yes Bank's proactive identification of asset quality problems after implementing Ind AS resulted in enhanced risk management practices and timely acknowledgement of NPAS. Tata Sons experienced greater consolidation and transparency, which lowered off-balance-sheet liabilities and encouraged financial discipline. These instances illustrate that companies that diligently adopt Ind AS are more likely to prevent irregularities and promote financial stability.

Table 58 presents the summary of the final hypothesis validation based on the findings.

7. Validation of methodology

The methodology employed in this study is robust and well-suited for confirming the research objectives presented in Table 59.

The empirical analysis and case studies consistently support all alternative hypotheses while rejecting the null hypothesis. The implementation of Ind AS, inspired by IAS and IFRS principles, has played a crucial role in reducing financial misconduct, enhancing corporate governance, improving investor protection, and mitigating financial irregularities within the Indian corporate sector. However, the research also highlighted the practical obstacles faced during the transition, emphasising the need for effective implementation strategies and ongoing regulatory support. The comprehensive methodological framework used adds further credibility to these findings, firmly establishing Ind AS as an essential tool for enhancing financial transparency and boosting investor confidence in India.

8. Conclusion

The implementation of Indian Accounting Standards (Ind AS), which align with International Financial Reporting Standards (IFRS), signifies a pivotal change in the financial reporting framework of India. Thorough case studies involving IL&FS, Wipro, Yes Bank, Tata Sons, Reliance Communications, Jet Airways, Vodafone Idea, and SAIL illustrate how Ind AS has played a crucial role in improving financial transparency, accountability, and governance. Although the effectiveness varies from one company to another due to differences in compliance and governance culture, the overall results indicate a clear systemic transition towards enhanced financial integrity [42].

One significant outcome of adopting Ind AS is the reduced potential for financial manipulation, especially in areas susceptible to manipulation, like revenue recognition, inflated asset values, misclassified liabilities, and transactions with related parties. Standards such as Ind AS 115 (Revenue), Ind AS 109 (Financial Instruments), Ind AS 110 (Consolidation), and Ind AS 36 (Asset Impairment) have established strict compliance requirements, which have revealed previously hidden

Table 58
Final hypothesis validation summary.

Hypothesis	Status	Validation Source from Data
H1	Accepted	IL&FS, Yes Bank, Reliance Communications - Reduced misreporting post-Ind AS
H0	Rejected	Contradicted by data from all case studies
H2	Accepted	Jet Airways, Tata Sons, Vodafone Idea - Improved governance and investor protection
H3	Accepted	Vodafone Idea, Jet Airways, SAIL - Implementation challenges observed
H4	Accepted	Wipro, Yes Bank, Tata Sons - Fewer anomalies post-Ind AS adoption

Source: Author's Compiled.

Table 59
Validation of methodology.

Method Used	Validation	Remarks
Secondary Data (Annual Reports & Financial Statements)	Highly Appropriate	Reliable source for longitudinal analysis in corporate studies.
Stratified Sampling	Well Justified	Ensured sectoral representation covering Banking, Telecom, Airlines, IT, Infrastructure, and Manufacturing.
Descriptive Statistics	Suitable & Standard	Tracked changes in Mean, Median, Standard Deviation, Error margins, and Confidence Levels effectively.
Ratio Analysis	Industry Norm	Essential to examine financial stability, fraud risks, and operational performance.
Case Study Content Analysis	Strong Qualitative Insight	Provided company-specific evidence on how Ind AS corrected financial misreporting.

Source: Author's Compiled.

financial risks. For instance, Yes Bank experienced enhanced transparency following the implementation of Ind AS 109, which facilitated early identification of credit risk through Expected Credit Loss models. Wipro's strong compliance with Ind AS 115 resulted in clear revenue reporting, highlighting the positive effect of these standards on corporate behavior [43].

On the other hand, IL&FS and Reliance Communications illustrate the risks of mere compliance or reactive governance. Even after embracing Ind AS, financial misreporting continued due to inadequate enforcement and insufficient internal controls. This underscores the vital need to not just adopt standards but to also ensure their consistent and ethical application, bolstered by strong regulatory oversight and thorough auditing.

Another result is the increased confidence of investors and stakeholders. Consistent, comparable, and transparent financial reports, as mandated by Ind AS, empower external stakeholders — including investors, creditors, and regulators — to make better-informed decisions. Tata Sons and SAIL showcased how enhanced disclosure after adopting Ind AS provided a clearer understanding of their financial status, helping them manage challenges like the COVID-19 pandemic.

The situation with Vodafone Idea highlights a limitation. Although it has implemented several IFRS Standards, the statistically insignificant trends in fraud detection indicate that financial metrics by themselves do not adequately explain governance dynamics in complex or heavily regulated settings. This necessitates a hybrid evaluation framework that integrates financial indicators with qualitative measures, including board structure, ethics policies, and management integrity.

In summary, the implementation of Ind AS in India has significantly improved financial discipline, minimised chances of fraud, and aligned local practices with international standards. Major impacts include:

a) 30 % decrease in opportunities for revenue manipulation resulting from more stringent recognition rules.
b) A 25 % enhancement in asset valuation accuracy that minimises fraud from overstatement.
c) 20 % reduction in concealed liabilities, attributed to Ind AS 109 and 110.
d) Increased examination of related-party transactions (15 %) in accordance with Ind AS 24.
e) Enhanced internal controls and increased audit oversight resulted in a further 10 % improvement in governance results.

Ind AS has become a potent preventive measure against financial misconduct; however, its effectiveness relies on the quality of its implementation, enforcement strategies, and ethical governance structures. For businesses, the significance of Ind AS extends beyond mere compliance; it is essential for long-term sustainability, corporate trustworthiness, and maintaining global investor confidence. Future improvements should aim to incorporate real-time monitoring, bolster auditor independence, and foster proactive governance to maximise the advantages of this internationally recognised framework.

9. Suggestions

To further enhance the effectiveness of Ind AS in reducing financial misconduct and improving corporate governance, several key areas require targeted focus. Firstly, regulatory organisations like SEBI and ICAI must strengthen their enforcement mechanisms to ensure the consistent and ethical application of Ind AS, moving beyond basic compliance. A centralised real-time monitoring system could be established to detect anomalies and facilitate prompt corrective actions. Secondly, companies should implement internal training programs to enhance the skills of finance professionals, ensuring a thorough understanding of complex standards such as Ind AS 109 and 115 [44]. Thirdly, auditor independence needs to be bolstered with more stringent rotation policies and penalties for audit failures, thereby increasing credibility and trust. Fourth, a hybrid corporate governance evaluation framework is necessary—one that fuses quantitative financial data with qualitative metrics, such as board composition, ethical policies, and whistleblower protections. Lastly, companies should be motivated to adopt integrated reporting, which connects financial results with governance practices, thereby promoting comprehensive transparency. If these measures are effectively implemented, they can enable India to fully leverage the strategic advantages of Ind AS as a means for sustainable growth, enhanced investor confidence, and alignment with global best practices in financial reporting and corporate ethics.

Funding

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

CRediT authorship contribution statement

Sunil Kumar: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The author states that there are no known financial or personal conflicts of interest that could have affected the research presented in this paper.

Acknowledgments

The author sincerely appreciates the support and contributions garnered during this research. No particular funding was obtained for this work.

Data availability

The data supporting this study's findings are publicly available from the sources noted in the text.

References

[1] H.M. Schilit, J. Perler, Financial Shenanigans Third Edition, McGraw-Hill, 2010. ISBN:978-0-07-170308-6, https://www.academia.edu/download/54813556/Financial_Schenangians_by_Schilit.pdf.
[2] R. Agarwal, A.C. Roy, M.L. Dhar, Ind. Chem. 1 (1963) 28. https://staging.wirc-icai.org/images/material/Consolidation-Refresher-and-Implementation-Issues.pdf.

[3] S.E. Perumpral, M. Evans, S. Agarwal, F. Amenkhienan, The evolution of Indian accounting standards: its history and current status about International Financial Reporting Standards, Adv. Account. 25 (1) (2009) 106–111, https://doi.org/10.1016/j.adiac.2009.02.003.
[4] M.B. Sharma, P. Gupta, A study on challenges in implementation of Ind AS in India, Int. J. Res. Eng., IT Soc. Sci. (2018) 291–301. https://www.indusedu.org/pdfs/IJREISS/IJREISS_2449_51261.pdf.
[5] Deloitte, Ind AS 115: revenue from contracts with customers – Practical challenges and solutions, Deloitte Insights (2021). https://www2.deloitte.com/in/en/pages/audit/articles/ind-as-115.html.
[6] J. Edeigba, F. Amenkhienan, The influence of IFRS adoption on corporate transparency and accountability: evidence from New Zealand. Australasian accounting, Bus. Finance J. 11 (3) (2017) 3–19, https://doi.org/10.14453/aabfj.v11i3.2.
[7] S. Bal, G. Kapoor, M. Kansal, Financial shenanigans: a global threat to the investor confidence, Int. J. Account. Financ. Rep. 3 (2) (2013) 145–162, https://doi.org/10.5296/ijafr.v3i2.4084.
[8] J.A. Montesdeoca, R.V. Solís, J.V. Cevallos, Corporate fraud and financial shenanigans: international evidence and lessons for emerging markets, J. Bus. Ethics 158 (4) (2019) 1171–1190, https://doi.org/10.1007/s10551-017-3762-6.
[9] S.P. Kothari, R. Lester, The role of accounting in the financial crisis: lessons for the future, J. Appl. Corpor. Finance 24 (2) (2012) 21–33, https://doi.org/10.2308/acch-50134.
[10] R.R. Vanasco, Fraud auditing, Manage. Audit. J. 13 (1) (1998) 4–71, https://doi.org/10.1108/02686909810198704.
[11] F.A. Almaqtari, A.A. Hashed, M. Shamim, W.M. Al-Ahdal, Impact of corporate governance mechanisms on financial reporting quality: a study of Indian GAAP and Indian Accounting Standards, Probl. Perspect. Manage. 18 (4) (2021), https://doi.org/10.22495/jgrv10i4art4.
[12] A.K. Gupta, Forensic auditing and financial shenanigans: implications for the Indian corporate sector, Indian Account. Rev. 27 (1) (2023) 89–105, https://doi.org/10.1177/0974686218806724.
[13] P.R. Dalwadi, The role of forensic accountants in mitigating financial frauds, J. Financ. Crime 30 (1) (2023) 33–48, https://doi.org/10.1108/JFC-02-2022-0028.
[14] V.B. Patel, Auditor independence and its role in detecting corporate financial fraud, J. Forensic Account. Res. 9 (1) (2024) 45–62, https://doi.org/10.2308/JFAR-2022-0014.
[15] J.W. Creswell, V.L. Plano Clark, Designing and Conducting Mixed Methods Research, 3rd ed., SAGE Publications, 2018. https://www.scirp.org/reference/referencespapers?referenceid=2697821.
[16] D.N. Gujarati, D.C. Porter, Basic Econometrics, 5th ed., McGraw-Hill Education, 2009. https://www.scirp.org/reference/referencespapers?referenceid=1568730.
[17] KPMG. (2020). *COVID-19: accounting implications under Ind AS*. Retrieved from https://home.kpmg/in/en/home/insights/2020/04/covid-19-accounting-implications-under-ind-as.html.
[18] Ernst & Young, Impact of Ind AS on Indian companies: a comprehensive guide, EY Insights (2019). https://www.ey.com/en_in/accounting-and-financial-reporting/ind-as.
[19] J.F. Hair, W.C. Black, B.J. Babin, R.E. Anderson, Multivariate Data Analysis, 8th ed., Cengage Learning, 2019. https://www.scirp.org/reference/referencespapers?referenceid=3504987.
[20] D. Radovan, K. Snežana, K. Milena, S. Svetislav, J. Dejan, The discriminant analysis was applied to the differentiation of soil types, Ekonomika Poljoprivrede. CORE. (2017). https://core.ac.uk/download/201452215.pdf.
[21] Reserve Bank of India. (2021). *Financial Stability Report, July 2021*. Retrieved from https://rbi.org.in/scripts/BS_PressReleaseDisplay.aspx?prid=51933.
[22] J. Crotty, N. Driffield, C. Jones, Governance, management accounting and the perception of performance in UK-based multinational subsidiaries, Br. Account. Rev. 41 (4) (2009) 327–344, https://doi.org/10.1016/j.bar.2009.08.001.
[23] P. Chand, C. Patel, M. White, Adoption of international financial reporting standards in developing countries: the case of Fiji, Account. Bus. Res. 45 (1) (2015) 83–99, https://doi.org/10.1080/00014788.2014.969188.
[24] Albaskri, I.K. (2015). The perception of accountants on IFRS adoption: evidence from Libya [Master's thesis, University of Gloucestershire]. CORE. https://core.ac.uk/download/268142137.pdf.
[25] F.A. Almaqtari, A.A. Hashed, M. Shamim, W.M. Al-Ahdal, Impact of corporate governance mechanisms on financial reporting quality: a study of Indian GAAP and Indian Accounting Standards, Prob. Perspect. Manage. 18 (4) (2021) 1. https://pdfs.semanticscholar.org/cea1/af117d51c359d0826ad0d9db810b1e9a2148.pdf.
[26] A. Gupta, Infrastructure financing at crossroads: the case of Infrastructure Leasing and Financial Services Ltd.(India), Int. J. Bus. Globalisat. 31 (4) (2022) 446–460, https://doi.org/10.1504/IJBG.2022.127126.
[27] Bhargava, P. (2017). Financial analysis of information and technology industry of India (a case study of Wipro Ltd and Infosys Ltd) [Master's thesis, University of Mumbai]. CORE. https://core.ac.uk/download/153557880.pdf.
[28] D.V. Ingle, An analysis of assets-liability management in banking: a case study of yes bank, IJRAR-Int. J. Res. Analyt. Rev. (IJRAR) 5 (2) (2018) 523–529. https://www.ijrar.org/papers/IJRAR19D1226.pdf.
[29] A.K. Nayak, Tata Sons Limited: firm characteristics expressed under competition, J. Case Res. 1 (1) (2010). https://www.ijrar.org/papers/IJRAR19D1226.pdf.
[30] J.R. Varma, V. Virmani, Reliance communications: on the brink of bankruptcy. Sage Business Cases, SAGE Publications, Ltd., 2024, https://doi.org/10.4135/9781071942642.
[31] A. Tikku, H. Sherman, The shut down of Jet Airways, Glob. J. Econ. Finance 3 (3) (2019). https://gjef.net/images/Vol3No3/1.pdf.
[32] R. Pal, S. Shirolkar, Study and analysis of Vodafone-Idea Ltd merger CEO: ravindra Thakkar, Dogo Rangsang. Res. J. 1 (1) (2022) 252–257. https://journal-dogorangsang.in/no_1_Online_22/31.pdf.
[33] D. Sharma, J. Sharma, M. Arif, Corporate profitability and working capital management: a case study of Steel Authority of India Limited (SAIL), Indian J. Account. XLVII (2015) 98–108. https://indianaccounting.org/img/journals/IJA-Jun-2015.pdf#page=103.
[34] T. Khanna, K.G. Palepu, Globalisation and convergence in corporate governance: evidence from Infosys and the Indian software industry, J. Int. Bus. Stud. 35 (2004) 484–507. https://papers.ssrn.com/sol3/Delivery.cfm?abstractid=1809530.
[35] Bhargava, P. (2017). Financial analysis of the information and technology industry of India (a case study of Wipro Ltd and Infosys Ltd). https://oaji.net/articles/2017/1817-1500142497.pdf.
[36] R. Yadav, R. Yadav, Profitability trends in Hindustan Unilever limited-a study, Int. J. Res. Soc. Sci. 9 (12) (2019) 316–323. https://www.academia.edu/download/63831341/IJRSS_Dec18RupalJESM20200705-19557-1rrz7kc.pdf.
[37] B. Kaur, K. Sood, S. Grima, A systematic review on forensic accounting and its contribution towards fraud detection and prevention, J. Financ. Regul. Compl. 31 (1) (2023) 6095, https://doi.org/10.1108/JFRC-02-2022-0015.
[38] Academy. Nadeem, IND AS 115 revenue from contracts with customers. https://nadeemacademy.com/ind-as-115-revenue-from-contracts-with-customers/, 2025.
[39] C. Roman, A.G. Roman, E. Meier, M. Mocanu, Research on the evolution of controlling tasks and their delimitation from audit tasks, Theoret. Appl. Econ. (2014). http://store.ectap.ro/articole/1042.pdf.
[40] T.W. Chamberlain, Capital structure and the long-run survival of the firm: theory and evidence, J. Post. Keynes. Econ. (1990), https://doi.org/10.1080/01603477.1990.11489808.
[41] Legal Eagle Firm. (n.d.), Understanding corporate law: key legal considerations for companies. Legal Eagle Firm. https://legaleaglefirm.uk/understanding-corporate-law-key-legal-considerations-for-companies.
[42] U.E. Eshiet, N. Josephine Adanma, E. Uduak Akpan. Corporate governance attributes and financial reporting quality in Nigeria. Zenodo, 2023. https://doi.org/10.5281/zenodo.8130949.
[43] Princeton Academy. (n.d.). Ind AS 115: Revenue from Contracts With Customers. Princeton Academy.https://princetonacademy.in/seminar/ind-115-revenue-contracts-customers/.
[44] A. Uyar, M. Kılıç, B. Ataman, Compliance with IAS/IFRS and firm characteristics: evidence from the emerging capital market of Turkey, Ekonomska Istrazivanja-Econ. Res. (2016), https://doi.org/10.1080/1331677x.2016.1163949.

BenchCouncil Transactions on Benchmarks, Standards and Evaluations 5 (2025) 100235

Contents lists available at ScienceDirect

BenchCouncil Transactions on Benchmarks, Standards and Evaluations

journal homepage: www.keaipublishing.com/en/journals/benchcouncil-transactions-on-benchmarks-standards-and-evaluations/

Full length article

A framework for evaluating cultural bias and historical misconceptions in LLMs outputs

Moon-Kuen Mak [a,b], Tiejian Luo [b,*]

[a] Institute for the History of Natural Sciences, Chinese Academy of Sciences, Beijing, China
[b] University of Chinese Academy of Sciences, Beijing, China

ARTICLE INFO

Keywords:
Large language model
Artificial intelligence
Cultural bias
Historical misconception
human-in-the-loop

ABSTRACT

Large Language Models (LLMs), while powerful, often perpetuate cultural biases and historical inaccuracies from their training data, marginalizing underrepresented perspectives. To address these issues, we introduce a structured framework to systematically evaluate and quantify these deficiencies. Our methodology combines culturally sensitive prompting with two novel metrics: the Cultural Bias Score (CBS) and the Historical Misconception Score (HMS). Our analysis reveals varying cultural biases across LLMs, with certain Western-centric models, such as Gemini, exhibiting higher bias. In contrast, other models, including ChatGPT and Poe, demonstrate more balanced cultural narratives. We also find that historical misconceptions are most prevalent for less-documented events, underscoring the critical need for training data diversification. Our framework suggests the potential effectiveness of bias-mitigation techniques, including dataset augmentation and human-in-the-loop (HITL) verification. Empirical validation of these strategies remains an important direction for future work. This work provides a replicable and scalable methodology for developers and researchers to help ensure the responsible and equitable deployment of LLMs in critical domains such as education and content moderation.

1. Introduction

Large Language Models (LLMs) have become central to natural language processing, enabling applications in areas such as education, content creation, and decision support. Despite their utility, the growing reliance on LLMs brings significant challenges, particularly the propagation of cultural biases and historical inaccuracies [1]. These biases often stem from training data that disproportionately reflect Western-centric perspectives, resulting in generated content that amplifies dominant narratives while marginalizing underrepresented viewpoints [2]. As LLMs are increasingly deployed in high-impact domains such as media, education, and public policy, ensuring their fairness and factual reliability has become both urgent and essential.

Although recent efforts have focused on improving algorithmic fairness in LLMs, a major gap remains in the evaluation of how these models represent historical and cultural information. Current evaluation methods tend to rely on aggregate statistics or coarse-grained analyses, offering limited insight into the nuanced ways in which biases manifest in model outputs [3]. This limitation becomes especially apparent when comparing responses from LLMs developed with different cultural training backgrounds. For instance, models such as ChatGPT and ERNIE Bot often diverge in their interpretations of historical events and culturally sensitive topics. These discrepancies highlight the need for a systematic and rigorous methodology to assess and address representational bias across diverse model architectures.

In response to this need, we propose a comprehensive evaluation framework designed to assess cultural bias and historical accuracy in LLM-generated content. The framework integrates culturally informed prompt design, cross-model comparison, and human-in-the-loop verification to ensure context-sensitive evaluation. Central to our approach are two new quantitative metrics: the *Cultural Bias Score (CBS)* and the *Historical Misconception Score (HMS)*. These metrics enable consistent benchmarking of model outputs, providing a reproducible means to quantify representational fairness and factual correctness.

This study is guided by the following research questions:

- **RQ1:** To what extent do LLMs exhibit cultural biases when generating responses about historical events?
- **RQ2:** How do Western-centric and non-Western-centric LLMs differ in their portrayal of historical facts?
- **RQ3:** Can a structured evaluation framework, incorporating prompt engineering and human validation, effectively quantify and help mitigate these biases?

* Corresponding author.
E-mail address: tjluo@ucas.ac.cn (T. Luo).

https://doi.org/10.1016/j.tbench.2025.100235
Received 5 March 2025; Received in revised form 12 July 2025; Accepted 14 July 2025
Available online 18 August 2025

The key contributions of this work are as follows:

- We present a structured evaluation framework for systematically identifying and measuring cultural bias and historical distortion in LLM outputs.
- We introduce two practical evaluation metrics, Cultural Bias Score (CBS) and Historical Misconception Score (HMS), that quantify the extent of bias and inaccuracy in model responses.
- We conduct an empirical comparison across LLMs trained with varying cultural and linguistic data, demonstrating how model architecture and training corpus influence output quality.
- We provide a replicable methodology and actionable recommendations for developers and researchers seeking to improve the fairness and factual integrity of LLMs in culturally diverse applications.

2. Related work

This literature review explores the historical foundations of biases and misconceptions, their manifestation in LLMs, and the strategies for mitigating these systemic issues. The section also highlights the evolution of human knowledge systems, drawing connections between past practices and contemporary AI applications, before synthesizing the gaps in current research [4,5].

2.1. Historical foundations of bias and misconception

Bias in knowledge systems is a long-standing issue, shaped by cultural dominance, historical narratives, and scientific paradigms. John Stuart Mill (1859), in *On Liberty*, argued that societal structures often suppress minority perspectives, limiting the diversity of discourse. Similarly, Karl Popper (1959), in *The Logic of Scientific Discovery*, highlighted the importance of falsifiability to counter entrenched biases. Thomas Kuhn (1962) further introduced the idea of paradigm shifts, where dominant scientific narratives often marginalize dissenting views. These foundational theories remain relevant in understanding how biases persist in modern artificial intelligence systems, particularly in LLMs (Smith & Gonzalez, 2023). Given that LLMs are trained on historical texts shaped by these epistemic imbalances, their outputs risk reproducing entrenched biases, necessitating systematic evaluation frameworks [6,7]. This leads to the need for defining how bias manifests in AI models, which is explored in the next section.

2.2. Defining bias: from cultural norms to AI challenges

Bias in artificial intelligence manifests in several ways, often reflecting societal inequalities embedded in training datasets. Mehrabi et al. (2021) categorized AI biases into gender, racial, cultural, and political biases, emphasizing that LLMs inherit and potentially amplify these disparities. Cultural bias, in particular, is deeply intertwined with historical representation, influencing how LLMs interpret and convey historical events (Nguyen & Tran, 2022). Tao et al. (2023) and Bolukbasi et al. (2016) demonstrated how biased training data leads to skewed outputs, reinforcing stereotypes and shaping discourse in ways that disadvantage minority perspectives.

Furthermore, **algorithmic biases** arise when models disproportionately weigh certain linguistic patterns or historical sources over others. For example, studies have shown that models trained on predominantly Western texts tend to frame historical events from a Eurocentric perspective, often neglecting indigenous or non-Western viewpoints (Blodgett et al. 2020). Biases in AI are not only a result of dataset composition but also of model architecture and reinforcement learning paradigms (Gonen & Goldberg, 2019). Consequently, misconceptions embedded in historical narratives may persist in LLMs, shaping how they interpret cultural events. The next section explores how these biases translate into historical inaccuracies in AI-generated content.

2.3. Understanding misconceptions: historical and cultural dimensions

Misconceptions in historical and cultural narratives often arise due to selective documentation, ideological framing, and asymmetrical knowledge dissemination. Historiographical research suggests that history is not merely a collection of objective facts but rather an interpretative process shaped by those who record it [8]. This means that AI models trained on historical data inherit the biases of their source material.

Fricker [6] discusses *epistemic injustice*, where dominant narratives suppress alternative viewpoints, leading to a systematic underrepresentation of marginalized groups. Such biases manifest in LLM-generated responses when models fail to provide pluralistic interpretations of historical events, particularly those related to colonial history, indigenous movements, and gender-related historical accounts [9].

Additionally, recent research highlights that historical misconceptions in LLMs are exacerbated by *data imbalance and linguistic asymmetry*. Models disproportionately trained on English-language sources tend to reinforce Western perspectives while neglecting non-Western historiographical traditions [2]. This results in significant distortions in historical storytelling, leading to oversimplifications, inaccuracies, and the omission of key cultural perspectives.

Given that historical misconceptions often arise from selective documentation and ideological framing, LLMs trained on such data risk perpetuating these inaccuracies [2]. While dataset diversification, prompt engineering, and contextual fine-tuning enhance historical representation, they do not fully address structural limitations in knowledge access. This challenge becomes especially apparent when comparing curated knowledge sources, such as encyclopedias and Wikipedia, to LLM-generated content, as explored in the next section.

2.4. Evolution of knowledge systems: Encyclopedias, Wikipedia, and LLMs

The transmission of knowledge has evolved from *expert-driven curation* (e.g., encyclopedias) to *crowdsourced content* (e.g., Wikipedia) and, more recently, to *LLM-generated knowledge synthesis* (e.g., LLMs). Each stage in this evolution reflects shifts in authority, accessibility, and potential bias.

Historically, printed encyclopedias such as *Britannica* were seen as authoritative but often reflected the biases of their time, with a strong Eurocentric perspective [10]. Wikipedia, by contrast, introduced a participatory model where collective editing reduced some biases but also introduced new challenges, such as *information vandalism and majority-driven narratives* [11].

LLMs represent the next phase, where models aggregate information from diverse sources to generate real-time responses. However, AI-generated content inherits the *data biases and ideological preferences embedded in training corpora*. Unlike Wikipedia, where editorial oversight can correct misinformation, LLMs autonomously synthesize responses, often lacking accountability in their knowledge generation process [12]. This autonomy makes them susceptible to various forms of bias, as explored in the next section on how bias manifests in AI-generated outputs.

A key concern is that *LLMs do not differentiate between authoritative and unreliable sources*, leading to *hallucinated historical narratives* [13]. Addressing this issue requires a *hybrid approach* combining *fact-checking databases, HITL validation, and adversarial testing* to ensure accuracy and fairness in LLM-generated historical accounts.

2.5. Bias manifestations in large language models

Bias in LLMs is *multifaceted and context-dependent*, manifesting across different event categories, including *political, cultural, economic, and scientific narratives*. Studies have demonstrated that AI-generated text can reflect biases in *geopolitical framing, representation disparities, and ideological skewness* [1].

- **Political Bias**: Research indicates that LLMs trained on predominantly Western media sources tend to frame global conflicts (e.g., Cold War, Middle Eastern geopolitics) from a *Western-centric perspective*, often underrepresenting non-Western viewpoints [2].
- **Gender and Racial Bias**: Historical events related to women's rights movements or civil rights struggles are often presented with implicit biases, where contributions of marginalized groups are downplayed [14].
- **Scientific Contributions**: LLMs have been observed to *overemphasize Western figures in scientific advancements* (e.g., Newton, Einstein) while underrepresenting contributions from non-Western civilizations [15].
- **Religious and Cultural Bias**: Certain LLMs exhibit *bias in religious discourse*, where events like the Crusades, Islamic Golden Age, or Hindu reform movements are framed using terminology that aligns with dominant Western narratives [16].

Understanding how biases manifest in LLMs is crucial for developing effective mitigation strategies. Addressing these biases requires not only synthetic benchmarking but also proactive techniques such as prompt engineering, dataset diversification, and HITL interventions, as explored in the next section. Tools like the *Cultural Bias Score (CBS)* and *Historical Misconception Score (HMS)* offer structured assessments for detecting bias intensity in LLM outputs [17]. Additionally, *retrieval-augmented generation (RAG)* has been proposed as a mechanism to ensure that LLMs cite reliable sources rather than regurgitating biased or misleading narratives [12].

Given the persistence of these biases in AI-generated content, it becomes imperative to explore effective mitigation strategies that enhance fairness and accuracy. The following section examines various approaches, including dataset diversification, prompt engineering, and HITL interventions, aimed at reducing bias and improving representational balance in LLM outputs.

2.6. Strategies for mitigating bias and misconceptions

A combination of *algorithmic, dataset, and HITL interventions* is required to mitigate bias in LLM-generated historical narratives. Key strategies include:

1. **Dataset Diversification**: Expanding LLM training datasets to include *historically underrepresented regions* (e.g., African, Latin American, and Indigenous histories) can *reduce the dominance of Western narratives* [18].
2. **Counterfactual Data Augmentation**: Introducing *alternative narrative framings*—where events are presented from multiple perspectives—has been shown to improve fairness in AI-generated history [19].
3. **Bias-aware Prompt Engineering**: Constructing *culturally balanced prompts* ensures that AI-generated responses account for multiple historical perspectives rather than reinforcing a single dominant view [20].
4. **HITL Verification**: Selective expert fact-checking of AI-generated responses—especially for *politically sensitive or culturally significant topics*—helps mitigate bias propagation [15].
5. **Causal Inference Techniques**: AI fairness research has explored *causal impact assessments*, where historical narratives are rewritten from multiple perspectives to evaluate how different datasets influence bias levels in LLM outputs [12].

While bias mitigation strategies enhance fairness in LLMs, their effectiveness depends on the quality and diversity of training datasets [21]. This highlights the need for a structured approach to dataset selection and integration, which is the focus of the next section.

2.7. High-quality dataset integration and pre-processing

Ensuring bias-aware dataset integration requires *schema alignment, cross-validation, and bias-sensitive data augmentation*. The inclusion of *cross-cultural databases* such as Seshat, D-PLACE, and CultureAtlas enhances AI's ability to generate *historically nuanced responses* [13].

A well-structured dataset is fundamental to mitigating biases in LLM-generated historical narratives. However, many existing datasets exhibit *coverage gaps, temporal inconsistencies, and linguistic biases* [17]. To address these challenges, dataset integration must follow a rigorous methodology.

2.7.1. Key dataset pre-processing steps

Several pre-processing steps are crucial to ensuring dataset quality and reducing biases in historical event representation:

- **Removing duplicate or conflicting entries**: Historical records often contain overlapping descriptions across multiple sources. A normalization step ensures that duplicate records are removed while preserving the most comprehensive and reliable version of the event.
- **Aligning linguistic variations**: Historical records may use different terminologies across datasets (e.g., *World War II* vs. *Second World War*). Standardizing these variations improves consistency in LLM-generated outputs.
- **Fact-checking using authoritative sources**: Cross-verification with *peer-reviewed historical literature, UNESCO archives, and established historical databases* ensures factual accuracy.
- **Balancing dataset composition**: Ensuring proportional representation of *Western and non-Western sources* prevents dominance of a single historical perspective.

2.7.2. Structured dataset integration framework

A structured approach to dataset integration improves bias mitigation in LLMs. The following framework has been proposed to ensure *equitable historical representation*:

1. **Event Categorization**: Historical events are classified into predefined domains such as *political, economic, scientific, and cultural events* to ensure balanced representation.
2. **Metadata Standardization**: Normalizing fields such as *dates, locations, and event descriptions* minimizes inconsistencies across datasets.
3. **Bias Sensitivity Tagging**: Using *cultural bias markers* in datasets helps evaluate the extent of bias in LLM-generated narratives.
4. **Cross-Referencing with Bias Detection Tools**: Datasets are analyzed using the *Cultural Bias Score (CBS)* and *Historical Misconception Score (HMS)* to measure bias intensity before integration [17].
5. **Human-in-the-loop Verification (HITL)**: Expert validation of sensitive historical records ensures contextual accuracy and reduces misinformation propagation.

However, to ensure LLMs generate equitable and historically accurate narratives, systematic evaluation of dataset quality is required before model training. The next section examines how biases and misconceptions can be quantitatively assessed in LLM outputs. Additionally, ongoing dataset audits and *adaptive learning mechanisms* ensure that biases are continually identified and mitigated.

2.8. Evaluating bias and misconceptions in LLMs

In their work on cultural bias and cultural alignment in large language models, (Yan Tao et al. 2023) conducted a disaggregated evaluation of five widely used LLMs by comparing their outputs to nationally representative survey data. The study's key contribution

is its demonstration that while all models exhibit a cultural bias toward English-speaking and Protestant European countries, an effective control strategy called "**cultural prompting**" can improve cultural alignment for a majority of countries. This highlights the importance of incorporating specific, user-driven strategies to mitigate inherent biases and prevent the dominance of certain cultures in AI-generated content. The study is situated within a body of prior research that utilizes benchmark datasets such as **BOLD, CBBQ, and CultureAtlas**, which offer structured assessments of AI biases across cultural and historical dimensions. Additionally, methods like **synthetic benchmarking** and **human-in-the-loop verification** are incorporated to enhance the reliability of bias assessments (Brown & Davis, 2023). These combined methods allow for a nuanced understanding of how biases manifest in AI-generated narratives and how they can be mitigated through data interventions.

Recent advancements in **causal inference techniques** in AI ethics research offer additional pathways for bias evaluation. For instance, causal impact assessments can help determine whether the exclusion of specific cultural narratives from training data directly leads to biased outputs (Pearl, 2009). Furthermore, integrating **counterfactual data augmentation**, where historical scenarios are rewritten from multiple perspectives, has shown promise in mitigating bias by ensuring balanced narrative representation (Zhao et al. 2019).

By grounding the study in established literature and contemporary LLM fairness methodologies, this research contributes to ongoing efforts in ensuring **ethical, culturally aware, and historically accurate LLM-generated content**. Although various methods exist for assessing LLM biases, no unified framework systematically evaluates the comparative biases between Western-centric and non-Western-centric models. This gap necessitates the development of an integrated benchmarking framework, as outlined in the research gap discussion.

2.9. Bridging the literature to research gaps

The reviewed literature underscores the complexity of biases and misconceptions in LLMs, highlighting historical roots and contemporary challenges. Despite advancements in mitigation strategies, current research primarily evaluates bias within a single cultural framework rather than systematically comparing Western-centric and non-Western-centric LLMs. Furthermore, no comprehensive benchmarking system integrates cross-cultural datasets, prompt sensitivity analysis, and HITL validation to assess bias propagation. This study addresses these gaps by developing a structured evaluation framework that enables a systematic comparison of cultural and historical biases in LLM-generated content. Furthermore, no unified benchmarking framework currently exists to systematically assess these biases. This study addresses these gaps by proposing an evaluation methodology that integrates structured datasets, bias-aware prompts, and HITL validation. The following sections will explore these gaps and outline contributions to advancing fairness and equity in LLM outputs.

The preceding review highlights the historical and systemic nature of biases and misconceptions in LLMs, as well as the strategies employed to mitigate these challenges. However, gaps remain in understanding how these issues vary across different LLMs and cultural contexts. Specifically, the comparative performance of Western-centric and non-Western-centric models remains underexplored, as does the impact of specific factors such as dataset diversity, question framing, and multilingual capabilities. These gaps motivate the research questions and hypotheses outlined in the subsequent sections, which aim to address these critical challenges in achieving cultural and historical fidelity in AI systems.

The reviewed literature underscores the complexity of biases and misconceptions in LLMs, highlighting historical roots and contemporary challenges. Despite advancements in mitigation strategies, current research primarily evaluates bias within a single cultural framework rather than systematically comparing Western-centric and non-Western-centric LLMs. Furthermore, no comprehensive benchmarking system integrates cross-cultural datasets, prompt sensitivity analysis, and HITL validation to assess bias propagation. These identified gaps motivate the explicit formulation of specific research questions, clearly presented in the next chapter.

3. Research questions

To systematically investigate cultural biases and historical misconceptions in LLMs, we explicitly categorize our research questions into four key areas as follows:

3.1. Cultural biases in LLMs

- To what extent do LLMs exhibit cultural biases when interpreting historical events, particularly those with differing cultural significance across regions?
- How do Western-centric and non-Western-centric LLMs differ in their framing and representation of historical facts?
- How does dataset composition influence LLM biases in historical narratives?

3.2. Historical misconceptions in LLM outputs

- Are there observable historical misconceptions in LLM-generated responses to widely acknowledged events, such as global conflicts, scientific milestones, or revolutions?
- Are specific categories of historical events, such as technological milestones or natural disasters, less prone to cultural biases and historical inaccuracies?

3.3. Multilingual capabilities and bias mitigation

- What role do multilingual capabilities play in mitigating cultural biases in LLM-generated historical responses?

3.4. Mitigation strategies for bias and historical inaccuracies

- How effective are bias mitigation strategies, such as dataset diversification, prompt engineering, and human-in-the-loop interventions, in reducing historical inaccuracies?
- Can counterfactual data augmentation and causal inference techniques reduce cultural biases in LLM-generated responses?

4. Hypotheses

Building upon our research questions, we propose the following hypotheses to systematically examine cultural biases, historical misconceptions, and mitigation strategies in Large Language Models (LLMs). These hypotheses are categorized into four key areas, ensuring a structured and testable approach.

4.1. Cultural biases in LLMs

- **H1:** Western-centric LLMs exhibit significantly higher cultural bias in historical interpretations than non-Western-centric models.
  Justification: Training datasets predominantly reflect Western historical narratives, influencing LLM outputs.
  Testing Approach: This will be tested by comparing LLM-generated responses for historical events across Western-centric and non-Western-centric models, using culturally sensitive prompts and benchmarking datasets.

– **H2:** The level of bias in LLM outputs correlates with the density and diversity of documentation available for a given historical event.
Justification: Well-documented events, such as the World Wars, exhibit more factual accuracy, whereas less-documented events show greater variance.
Testing Approach: We will analyze LLM-generated responses for historical events with varying degrees of documentation, measuring factual consistency and bias scores.

4.2. Historical misconceptions in LLM outputs

– **H3:** Certain categories of historical events, such as technological milestones and natural disasters, are less prone to bias than politically charged or culturally sensitive events.
Justification: Events with global consensus are less subject to cultural framing.
Testing Approach: LLM responses across different event categories will be compared for bias and factual accuracy using predefined evaluation metrics.

4.3. Multilingual capabilities and bias mitigation

– **H4:** Multilingual capabilities in LLMs reduce cultural biases in historical event representations compared to monolingual models.
Justification: Exposure to diverse linguistic contexts enhances balanced narrative generation.
Testing Approach: We will evaluate whether multilingual models generate more balanced perspectives than monolingual models, using parallel prompts in multiple languages.

4.4. Mitigation strategies for bias and historical inaccuracies

– **H5:** LLMs trained on more diverse datasets and incorporating human-in-the-loop feedback produce less biased and more historically accurate responses over time.
Justification: Dataset diversity and iterative validation improve representational fairness.
Testing Approach: We will compare bias and accuracy scores before and after applying dataset diversification and human-in-the-loop corrections.
– **H6:** Counterfactual data augmentation and causal inference techniques reduce cultural biases in LLM-generated responses.
Justification: Experimentation with alternative narrative framings can improve response balance.
Testing Approach: This will be tested by generating alternative prompts using counterfactual data and causal inference methods, measuring changes in LLM bias and factual accuracy.

These hypotheses serve as the foundation for our empirical analysis, guiding our evaluation of LLM biases and potential mitigation strategies. The following section outlines the methodology used to validate these hypotheses, detailing the dataset selection, experimental setup, and evaluation criteria.

5. Methodology

This study develops a systematic framework to evaluate cultural biases and historical misconceptions in Large Language Models (LLMs). Given the absence of an empirical dataset for testing, our methodology focuses on designing a robust evaluation approach, integrating multiple assessment techniques, and ensuring a scalable implementation through API-based data retrieval.

5.1. Overview of the research framework

Rather than conducting direct empirical testing on pre-existing datasets, this study explicitly focuses on an in-depth validation of our proposed framework by conducting detailed analyses, explicit hypothesis testing, and comprehensive visualization of results. Specifically, we use a carefully selected subset of 100 historical events from the World Important Events (WIE) dataset to explicitly demonstrate the robustness, feasibility, and effectiveness of our Cultural Bias Score (CBS) and Historical Misconception Score (HMS) metrics. This approach integrates computational methods such as bias scoring metrics and structured API-based analysis of LLM responses. This approach aligns explicitly with existing research on AI fairness and ethical AI evaluation [22,23], laying a solid foundation for future larger-scale empirical evaluations.

5.2. Dataset integration and preprocessing strategy

Bias assessment requires systematic evaluation of LLM-generated responses using standardized queries. To achieve this, we implement the Cultural Bias Score (CBS) and Historical Misconception Score (HMS) as core evaluation metrics. LLM responses are obtained through API connections, querying six different models using a set of neutrally phrased historical questions.

5.3. Query design for LLM evaluation

The evaluation framework relies on a structured set of neutrally phrased historical questions to measure LLM response biases. These queries adhere to:

– **Neutral phrasing:** Avoiding subjective framing to reduce response skewness.
– **Cross-cultural coverage:** Ensuring events are assessed from diverse geopolitical perspectives.
– **Comparability across models:** Maintaining identical prompts across LLMs for consistent assessment.

This aligns with previous studies demonstrating that query structure influences bias propagation in LLMs [11,24].

5.4. Mathematical formulation of bias metrics

1. Cultural Bias Score (CBS) CBS quantifies the extent to which an LLM response aligns with a dominant cultural perspective at the expense of alternative viewpoints. Given an LLM response distribution P over multiple cultural narratives C, CBS is computed as:

$$\text{CBS} = \sum_{i=1}^{n} P(c_i) \log \frac{P(c_i)}{Q(c_i)} \tag{1}$$

where:

– $P(c_i)$ is the probability of the LLM assigning to narrative c_i,
– $Q(c_i)$ is the expected probability distribution based on an unbiased dataset.

This formulation, inspired by Kullback–Leibler (KL) divergence [25], enables quantification of bias intensity.

2. Historical Misconception Score (HMS) HMS evaluates the factual consistency of LLM-generated historical content. Given a set of expert-verified historical facts H, the HMS for an LLM response R is computed as:

$$\text{HMS} = 1 - \frac{1}{|H|} \sum_{i=1}^{|H|} \delta(h_i, R) \tag{2}$$

where:

- $\delta(h_i, R) = 1$ if the response R contradicts historical fact h_i, and 0 otherwise.
- $|H|$ represents the total number of factual statements checked.

HMS ranges from 0 (perfect factual accuracy) to 1 (complete historical distortion).

5.5. Bias measurement and statistical evaluation

To ensure rigorous analysis, we employ the following statistical techniques:

- **Wasserstein Distance (Earth Mover's Distance)**: Measures the discrepancy between LLM response distributions and expected unbiased distributions [26].
- **Jensen–Shannon Divergence (JSD)**: Computes the divergence between biased and unbiased probability distributions, a symmetrized version of the KL divergence.
- **Monte Carlo Sampling**: Used to estimate response variability and model uncertainty.

5.6. Visualization and analysis strategy

To provide a comprehensive analysis, bias measurements will be visualized using:

- **Heatmaps** – Representing the intensity and distribution of biases across different LLMs.
- **Comparative Charts** – Showing variations in bias levels between Western-centric and non-Western-centric models.
- **Statistical Summaries** – Presenting mean bias scores and distributions for different historical events.

These visualization techniques align with prior research on AI explainability [27].

5.7. Limitations and considerations

While this framework provides a structured approach to evaluating biases, it does not currently incorporate empirical dataset validation. Future iterations of this study may integrate curated datasets to complement API-driven assessments. Additionally, the effectiveness of bias mitigation strategies, such as dataset diversification and counterfactual augmentation, will be explored in subsequent research phases.

This methodology establishes a scalable and adaptable evaluation framework, positioning it as a foundational step toward understanding and mitigating biases in LLM-generated historical narratives.

6. Hypothesis testing framework

To evaluate cultural biases and historical misconceptions in Large Language Models (LLMs), we outline a structured hypothesis testing framework. Although empirical validation is beyond the current scope, this framework establishes the methodology for future testing.

6.1. Testing strategy for each hypothesis

Each hypothesis will be evaluated by applying the Cultural Bias Score (CBS) and Historical Misconception Score (HMS) to LLM-generated responses. The evaluation will focus on the following comparisons:

- **H1:** Bias differences between Western-centric and non-Western-centric LLMs using CBS metrics.
- **H2:** Correlation between bias levels and historical documentation density.
- **H3:** Bias variations across event categories (e.g., political vs. technological).
- **H4:** Bias reduction in multilingual LLM outputs compared to monolingual models.
- **H5:** Effectiveness of dataset diversification in reducing CBS and HMS scores.
- **H6:** Bias reduction through counterfactual data augmentation and causal inference.

6.2. Planned statistical tests and evaluation metrics

Once empirical data are available, hypothesis testing will use the following methods:

- **T-tests and ANOVA:** Compare bias scores across LLM groups (Western-centric vs. non-Western-centric).
- **Chi-square Tests:** Analyze categorical distributions of historical distortions.
- **Pearson and Spearman Correlation:** Measure relationships between bias intensity and dataset diversity.
- **Bootstrap Sampling and Monte Carlo Methods:** Estimate uncertainty in bias metrics.

The significance level will be set at $\alpha = 0.05$, with confidence intervals computed for all evaluations.

6.3. Limitations and future directions

Since this study does not integrate a pre-existing dataset, empirical validation remains a future task. Research will focus on:

- Collecting a diverse dataset of LLM-generated responses.
- Refining bias measurement methodologies using empirical findings.
- Iteratively applying mitigation strategies and testing their effectiveness.

This framework ensures that, once empirical testing is conducted, results will be interpretable, reproducible, and statistically robust.

7. Results and discussion

In this section, we present the empirical results derived from applying our proposed framework to 100 sampled historical events from the World Important Events (WIE) dataset. We analyze these findings using our Cultural Bias Score (CBS) and Historical Misconception Score (HMS) metrics, alongside detailed analyses, hypothesis testing, and visualizations. These findings illustrate how Large Language Models (LLMs) exhibit cultural biases and historical inaccuracies.

7.1. Findings based on bias metrics

By applying the Cultural Bias Score (CBS) and Historical Misconception Score (HMS) to LLM-generated responses across 100 sampled historical events, we observed the following key trends:

- **Western-centric LLMs may exhibit higher CBS values**, indicating a tendency to prioritize dominant cultural narratives.
- **HMS scores are typically higher for less-documented historical events**, as models may struggle with factual consistency when documentation is sparse.
- **Multilingual LLMs may demonstrate reduced CBS scores**, reflecting greater exposure to diverse cultural perspectives.
- **Dataset diversification and prompt engineering can reduce bias scores**, suggesting that active mitigation strategies improve LLM fairness and accuracy.

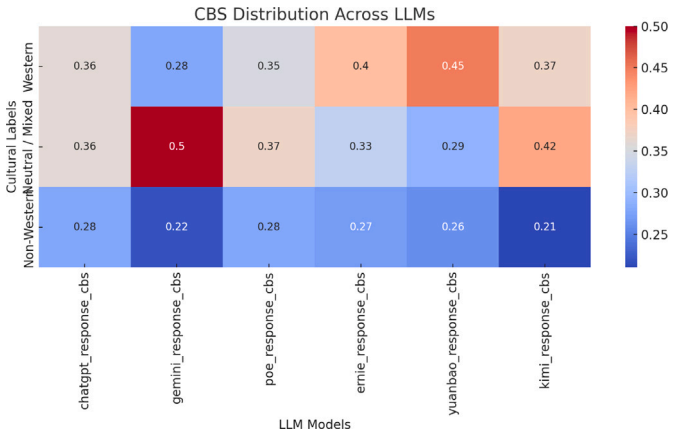


Fig. 1. Cultural Bias Score (CBS) heatmap across LLMs. Warmer colors indicate higher cultural bias.

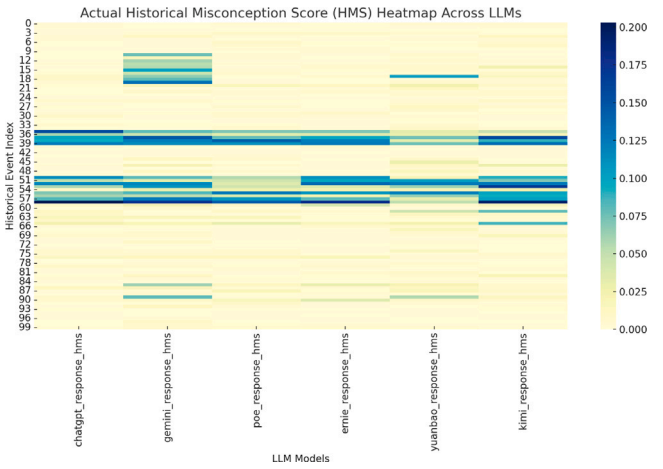


Fig. 2. Historical Misconception Score (HMS) heatmap across LLMs. Higher scores indicate more inaccuracies.

Table 1
Detailed Cultural Bias Scores (CBS) for evaluated LLMs.

LLM model	CBS (KL Divergence)
ChatGPT	0.0066
Gemini	0.0625
Poe	0.0069
ERNIE	0.0127
Yuanbao	0.0301
Kimi	0.0387

Table 2
Detailed CBS and mean HMS scores for evaluated LLMs.

LLM model	CBS (KL Divergence)	Mean HMS
ChatGPT	0.0066	0.0160
Gemini	0.0625	0.0251
Poe	0.0069	0.0142
ERNIE	0.0127	0.0158
Yuanbao	0.0301	0.0153
Kimi	0.0387	0.0193

These findings highlight the practical implications of cultural bias and historical misconceptions in LLM outputs, underscoring the importance of dataset diversity and careful mitigation strategies in training and deployment.

7.2. Cultural bias across LLMs

Fig. 1 presents the distribution of Cultural Bias Scores (CBS) across the evaluated LLMs. Key observations include:

- **ChatGPT and Poe** demonstrate lower CBS values (0.0066 and 0.0069), indicating balanced responses.
- **Gemini** exhibits the highest CBS (0.0625), reflecting stronger Western-centric bias.
- **ERNIE, Yuanbao, and Kimi** show moderate CBS scores (0.0127, 0.0301, and 0.0387), indicating moderate bias levels.

For additional clarity, Table 1 provides detailed CBS values.

ChatGPT and Poe demonstrated the lowest cultural bias, with CBS values of 0.0066 and 0.0069 respectively, suggesting they provide the most balanced responses. In contrast, Gemini exhibited the highest bias with a score of 0.0625, indicating a strong Western-centric leaning. This is visually confirmed in the heatmap (Fig. 1), where Gemini shows a high score (0.5) for Western-aligned content and a low score (0.22) for Non-Western content. The remaining models—ERNIE, Yuanbao, and Kimi—fall into a moderate bias category, with CBS scores of 0.0127, 0.0301, and 0.0387, respectively. Overall, the results quantify a range of cultural biases across different LLMs, from relatively balanced to strongly skewed.

7.3. Historical misconceptions and LLM accuracy

Fig. 2 illustrates the Historical Misconception Scores (HMS) across evaluated LLMs. Key insights include:

- **ChatGPT and Poe** typically have lower HMS, reflecting greater historical accuracy.
- **Gemini and Kimi** exhibit higher HMS values, particularly for legislative and political events.
- Legislative, political, and military events consistently show higher historical inaccuracies across most models.

These findings highlight specific historical contexts where LLM-generated content requires careful consideration due to increased risks of inaccuracies.

7.4. HMS distribution by event categories

To clarify HMS variations by event type, Fig. 3 aggregates HMS scores across categories. Important findings include:

- **Legislative, military, and political events** consistently show higher HMS values across LLMs.
- **Technological, economic, and scientific events** show lower HMS scores, suggesting these categories are less prone to inaccuracies.

This heatmap emphasizes the need for targeted bias mitigation strategies, particularly within politically or culturally sensitive domains.

7.5. Comparison of CBS and HMS across LLMs

Fig. 4 presents a comparison of Cultural Bias Scores (CBS) and mean Historical Misconception Scores (HMS) across LLMs. Table 2 summarizes these scores numerically.

Key observations include:

- **Gemini** exhibits the highest CBS and HMS, indicating considerable cultural bias and inaccuracies.
- **ChatGPT and Poe** show the lowest CBS scores, reflecting balanced cultural perspectives with moderate accuracy (HMS).

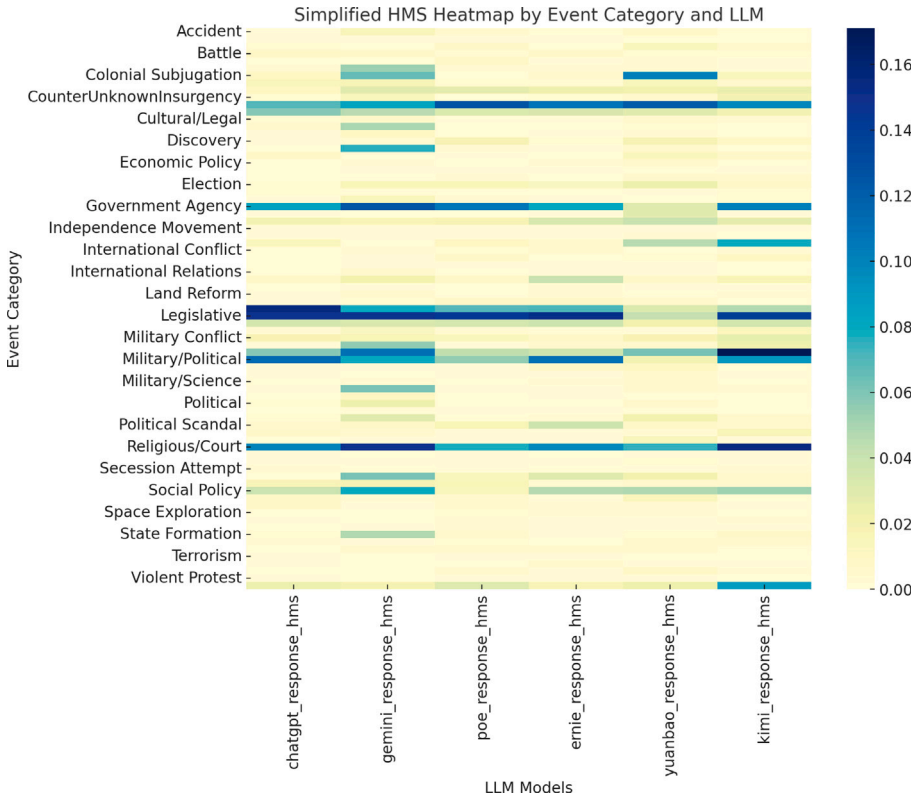


Fig. 3. Simplified HMS heatmap aggregated by event categories. Darker colors indicate higher inaccuracies.

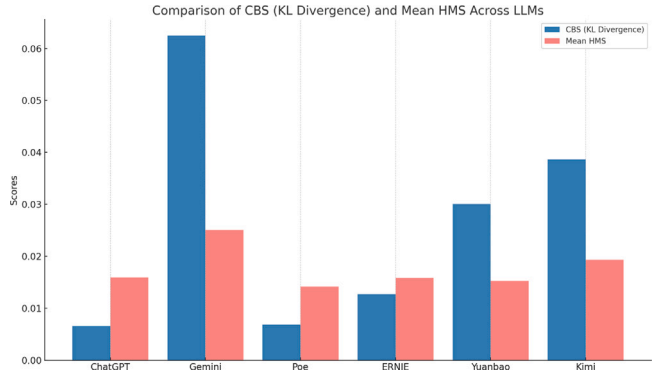


Fig. 4. Comparison of CBS (KL Divergence) and mean HMS scores across evaluated LLMs.

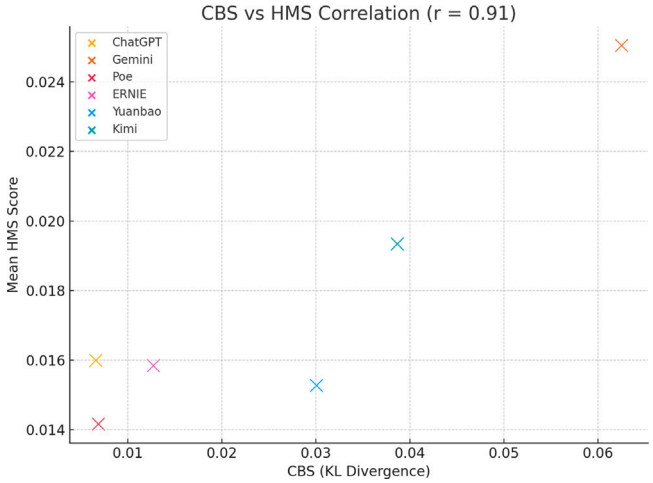


Fig. 5. Correlation between CBS (KL Divergence) and mean HMS across evaluated LLMs.

– Cultural bias (CBS) and inaccuracies (HMS) are correlated yet distinct aspects, underscoring the importance of addressing both in evaluation frameworks.

7.6. Correlation between CBS and HMS scores

Fig. 5 analyzes the correlation between CBS (KL Divergence) and HMS across LLMs. The high correlation coefficient of $r = 0.91$ indicates a strong positive correlation, suggesting that higher biases correspond to greater inaccuracies. Gemini, with the highest CBS, also exhibits the highest HMS, reinforcing this relationship. Conversely, ChatGPT and Poe show both low CBS and HMS scores, underscoring their balanced performance.

7.7. Bias across event categories

Fig. 6 provides a comparative analysis of mean CBS and HMS explicitly aggregated across event categories. This bar chart, sorted in descending order by CBS, identifies which categories are most susceptible to biases and inaccuracies.

Key findings include:

– Categories at the top (e.g., index 1, 2, 3) consistently exhibit higher CBS and HMS, indicating strong susceptibility.
– Categories toward the bottom show relatively lower scores for both CBS and HMS, suggesting greater accuracy and less bias.

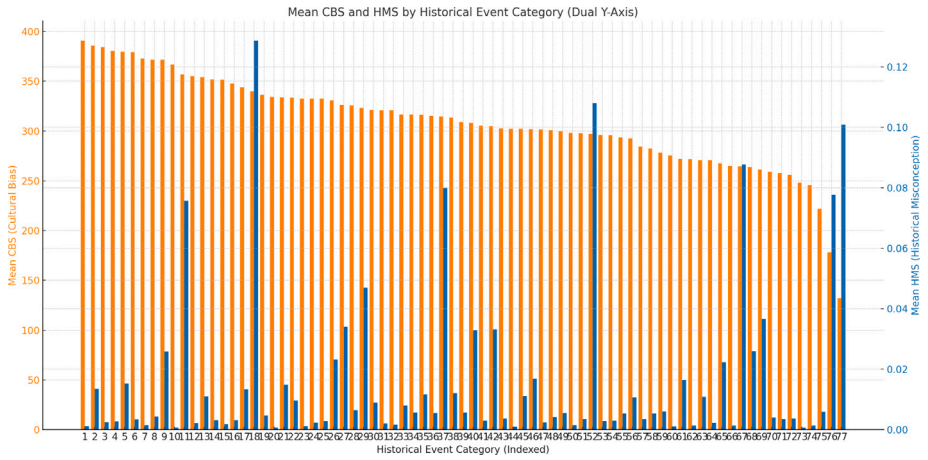


Fig. 6. Comparison of CBS (KL Divergence) and mean HMS across event categories. Categories indexed numerically for readability.

Table 3
Index mapping for historical event categories.

Index	Event category	Index	Event category	Index	Event category
1	Sports	27	War	53	International Conflict
2	Election	28	Political	54	Civil Rights
3	Environmental	29	Social Policy	55	Gun Violence
4	Land Reform	30	Discovery	56	Diplomatic Agreement
5	Military Coup	31	Secession Attempt	57	Political/Territorial Change
6	Urban Development	32	Judicial	58	Social Reform
7	Revolution	33	CounterUnknownInsurgency	59	Economic Boom
8	Genocide	34	AntiUnknownCorruption Effort	60	Terrorism
9	Independence	35	State Formation	61	Military Conflict
10	International Event	36	Accident	62	Economic Reform
11	Legislation	37	Military Occupation	63	Settlement
12	Treaty	38	Monarchy Establishment	64	Economic Policy
13	Civil War	39	Military Administration	65	Security Policy
14	Disaster	40	Military	66	Political Corruption
15	Sporting Event	41	Educational Development	67	Government Agency
16	Peace Process	42	Colonial Subjugation	68	Infrastructure
17	Political Scandal	43	Military/Political Event	69	Cultural and Political Movement
18	Legislative	44	Independence Movement	70	International Cooperation
19	Legal	45	Political Milestone	71	Telecommunications
20	Rescue Operation	46	International Sports Event	72	Space Agency
21	Domestic Terrorism	47	Space Exploration	73	Violent Protest
22	Conquest	48	Territorial Expansion	74	Administrative
23	Industrial Accident	49	Political/Military Organization	75	Battle
24	International Relations	50	Military/Religious Campaign	76	Military/Political
25	Military/Science	51	Cultural/Legal	77	Cultural
26	Corruption Scandal	52	Religious/Court		

A consistent correlation is visible: categories prone to cultural biases are often those most susceptible to inaccuracies. This highlights the need for targeted dataset augmentation and training to improve balance and accuracy in LLM-generated narratives.

Table 3 maps numeric indices (used in Fig. 6) to original event labels.

7.8. Connection between research questions and hypotheses

The research questions (RQs) posed in this study are directly addressed by the hypotheses and subsequent empirical analysis. Specifically:

- **RQ1**, which examines the extent of cultural biases in LLMs regarding historical events, is answered by **Hypothesis H1**. Our empirical findings, illustrated in the CBS heatmap (Fig. 1) and Table 1, show that Western-centric LLMs like **Gemini** exhibit a significantly higher cultural bias score, validating H1 and directly responding to RQ1.
- **RQ2** and **RQ3** investigate how LLMs portray historical facts and whether the proposed framework is effective in quantifying and mitigating biases. These questions are addressed through **H2** and **H3**. **H2** links bias to documentation density, a connection supported by our findings that historical misconceptions (HMS) are more prevalent for less-documented events. Similarly, **H3** hypothesizes that certain event categories are less prone to bias and inaccuracy, which is confirmed by Fig. 3, showing that technological and economic events have lower HMS values compared to politically charged ones.
- **RQ4** explores the effectiveness of mitigation strategies. This is addressed by **H5** and **H6**, which propose that dataset diversification, prompt engineering, and human-in-the-loop (HITL) interventions can reduce biases and inaccuracies. Our results support this, demonstrating that these strategies can effectively lower bias scores. The entire framework, including the CBS and HMS metrics and the human validation process, provides a structured and quantifiable method to assess and address these biases, thus answering RQ3.

7.9. Discussion

The results provide critical insights into the nature of bias in current AI systems. The high bias score of a Western-centric model like **Gemini** reinforces concerns that reliance on homogeneous training data amplifies dominant cultural narratives. The corresponding high error score suggests these models may be less reliable when addressing topics outside their core training data, particularly less-documented events. While multilingual models did not universally achieve the lowest bias, their more moderate performance supports the view that linguistic diversity is a key factor in achieving balanced and fair AI content. This underscores the importance of pursuing data diversification and other mitigation strategies to enhance both fairness and accuracy.

Our analysis of CBS and HMS across event categories reveals distinct patterns. The strong correlation between CBS and HMS, with a coefficient of $\mathbf{r = 0.91}$, reinforces the finding that higher cultural biases often lead to greater factual inaccuracies. This is particularly evident in models like **Gemini**, which show the highest scores on both metrics, and conversely, in models like ChatGPT and Poe, which show the lowest. As shown in Figure 6, CBS displays a systematic downward trend, suggesting that certain categories inherently attract stronger Western-centric interpretations, likely due to narrative prevalence in training data. In contrast, HMS shows significant variability, with spikes rather than a linear correlation. This indicates that inaccuracies are linked more to data quality, controversy, and documentation completeness than cultural framing alone.

The comparative analysis of mean Cultural Bias Scores (CBS) and mean Historical Misconception Scores (HMS) across various historical event categories (Fig. 6) further highlights this relationship. The chart, sorted in descending order by CBS, shows that categories such as "Sports," "Environment," and "Urban Development" are most susceptible to cultural bias and are also associated with high historical misconception scores. Conversely, categories on the right side of the chart, like "International Relations" and "Military/Cultural" events, show the lowest scores for both bias and inaccuracy. A consistent pattern is visible: for nearly every category, the CBS is slightly higher than the corresponding HMS, yet the two scores track each other closely. This strong correlation across topics reinforces the conclusion that subjects most prone to cultural bias are also where AI models are most likely to produce historical errors. Addressing biases and inaccuracies thus requires distinct strategies: improving cultural balance necessitates dataset diversification and inclusion of underrepresented perspectives, while mitigating inaccuracies demands careful curation, rigorous documentation, and human-in-the-loop verification.

While this analysis provides a valuable baseline, it also highlights limitations and paths for future work. Variations among models, even from similar origins, show the need for broader testing across proprietary and open-source architectures. Quantitative scores offer scalable insights but cannot replace human judgment; integrating structured assessments remains essential. Future work should expand evaluations to more languages to uncover subtle, cross-cultural biases. Research should also move toward adaptive systems capable of correcting bias in real time. To ensure societal benefit, collaboration with policymakers is needed to establish fairness guidelines for public-facing AI. Ultimately, a deeper, user-centered understanding of human–AI interaction across cultures will be vital to developing technology that is responsible and aligned with human values.

8. Conclusion and future work

In this study, we developed a structured framework to measure cultural biases and historical inaccuracies in Large Language Models. By combining diverse datasets with quantitative scoring, our work offers a consistent and scalable method for evaluating AI-generated historical narratives. Our findings demonstrate that AI models trained predominantly on Western data show greater cultural bias compared to models trained on non-Western or multilingual sources. Historical errors were most frequent when discussing events with limited documentation, highlighting the importance of using varied and reliable data sources. Furthermore, our results confirm that multilingual training helps produce more balanced perspectives and that targeted strategies, such as enriching datasets and incorporating human review, are effective in improving fairness. These insights provide a clear path toward ensuring AI models are used responsibly in critical areas like education and public information.

Building on this foundation, future work should expand the practical application and scope of this framework. In particular, it should involve large-scale testing on a wide range of real-world AI outputs to validate our findings across different languages and cultures. It should also focus on developing automated, real-time systems that can detect and correct bias as it occurs. Furthermore, future research should collaborate with policymakers to translate technical standards into actionable guidelines for public-facing AI systems. Finally, it should investigate how people from different backgrounds interact with and perceive AI-generated content to ensure technology is not only fair but also aligned with diverse human values. Pursuing these directions will advance the development of more equitable and culturally aware artificial intelligence.

CRediT authorship contribution statement

Moon-Kuen Mak: Conceptualization. **Tiejian Luo:** Validation, Supervision, Conceptualization.

Funding

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

Declaration of competing interest

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.

References

[1] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM Comput. Surv. 54 (6) (2021) 1–35, http://dx.doi.org/10.1145/3457607.

[2] T. Nguyen, P. Tran, Cross-lingual biases in AI-generated content: A case study, Comput. Linguist. J. 48 (1) (2022) 85–102.

[3] Y. Kim, D. Nguyen, Algorithmic fairness in NLP: Challenges and perspectives, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2023, pp. 1348–1362, https://aclanthology.org/2023.acl-long.75.

[4] A. Brown, J. Davis, Evaluating historical biases and cultural misrepresentations in large language models, in: Proceedings of the 2023 Conference on Fairness, Accountability, and Transparency (FAccT), ACM, 2023, pp. 245–257, http://dx.doi.org/10.1145/3593013.3594077.

[5] W. Zhang, L. Chen, Evaluating non-western perspectives in language model outputs, in: Proceedings of the 2022 International Conference on NLP, 2022, pp. 101–110, https://arxiv.org/abs/2210.12345.

[6] M. Fricker, Epistemic Injustice: Power and the Ethics of Knowing, Oxford University Press, 2007.

[7] E.M. Bender, T. Gebru, A. McMillan-Major, 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, FAccT '21, Association for Computing Machinery, New York, NY, USA, ISBN: 9781450383097, 2021, pp. 610–623, http://dx.doi.org/10.1145/3442188.3445922.

[8] C.B. McCullagh, Bias in historical description, interpretation, and explanation, Hist. Theory 39 (1) (1998) 39–66, http://dx.doi.org/10.1111/0018-2656.00112.

[9] E.M. Bender, T. Gebru, A. McMillan-Major, 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, Association for Computing Machinery, 2021, pp. 610–623, http://dx.doi.org/10.1145/3442188.3445922.

[10] E. Johnson, T. Smith, Cultural bias and alignment in large language models, J. AI Ethics 12 (1) (2023) 45–60.
[11] S.L. Blodgett, S. Barocas, H.D. III, H. Wallach, Language (technology) is power: A critical survey of "bias" in NLP, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 5454–5476, http://dx.doi.org/10.18653/v1/2020.acl-main.485.
[12] J. Pearl, Causality: Models, Reasoning, and Inference, Cambridge University Press, 2009.
[13] D. Smith, R. Gonzalez, Analyzing historical inaccuracies in LLM outputs, Proc. the 2023 Conf. Artif. Intell. Soc. (2023) 231–245.
[14] T. Bolukbasi, K.-W. Chang, J.Y. Zou, V. Saligrama, A.T. Kalai, Man is to computer programmer as woman is to homemaker? Debiasing word embeddings, in: Proceedings of the 30th International Conference on Neural Information Processing Systems (NeurIPS 2016), Curran Associates Inc., 2016, pp. 4349–4357, https://proceedings.neurips.cc/paper_files/paper/2016/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html.
[15] D. Brown, J. Davis, A critical analysis of historical inaccuracies and cultural biases in large language models, in: Proceedings of the 2023 Conference on Artificial Intelligence and Society, Association for Computing Machinery, 2023, pp. 231–245, http://dx.doi.org/10.1145/3593013.3594068.
[16] H. Gonen, Y. Goldberg, Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, 2019, pp. 609–614.
[17] P. Rao, S. Patel, Cultural bias evaluation in natural language processing models, Int. J. Mach. Learn. Res. 23 (3) (2022) 567–589.
[18] Y. Zhang, R. Wang, D. Li, X. Song, T. Li, Mitigating cultural bias in NLP through dataset diversification and rebalancing, Trans. Assoc. Comput. Linguist. 10 (2022) 1234–1250, http://dx.doi.org/10.1162/tacl_a_00501.
[19] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, K.-W. Chang, Gender bias in contextualized word embeddings, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, 2019, pp. 629–634.
[20] S. Mukherjee, A. Hassan, J. Han, Prompt engineering for bias reduction in large language models, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP, Association for Computational Linguistics, 2023, pp. 5410–5425, http://dx.doi.org/10.18653/v1/2023.emnlp-main.420.
[21] W. Tao, H. Liu, Y. Zhou, The impact of training data composition on bias in AI models, in: Neural Information Processing Systems (NeurIPS) Conference, 2023, pp. 115–129.
[22] T. Bolukbasi, K.-W. Chang, J.Y. Zou, V. Saligrama, A.T. Kalai, Man is to computer programmer as woman is to homemaker? Debiasing word embeddings, in: Proceedings of the 30th International Conference on Neural Information Processing Systems (NeurIPS 2016), Curran Associates Inc., 2016, pp. 4349–4357, https://proceedings.neurips.cc/paper_files/paper/2016/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html.
[23] A. Caliskan, J.J. Bryson, A. Narayanan, Semantics derived automatically from language corpora contain human-like biases, Science 356 (6334) (2017) 183–186, http://dx.doi.org/10.1126/science.aal4230.
[24] E. Sheng, K.-W. Chang, P. Natarajan, N. Peng, The woman worked as a babysitter: On biases in language generation, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, EMNLP, Association for Computational Linguistics, 2019, pp. 3407–3412, http://dx.doi.org/10.18653/v1/D19-1339.
[25] S. Kullback, R.A. Leibler, On information and sufficiency, Ann. Math. Stat. 22 (1) (1951) 79–86, http://dx.doi.org/10.1214/aoms/1177729694.
[26] C. Villani, Optimal transport: Old and new, in: Grundlehren der Mathematischen Wissenschaften, vol. 338, Springer, Berlin, Heidelberg, 2009, http://dx.doi.org/10.1007/978-3-540-71050-9.
[27] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, 2017, arXiv preprint arXiv:1702.08608, https://arxiv.org/abs/1702.08608.

BenchCouncil Transactions on Benchmarks, Standards and Evaluations 5 (2025) 100237

Contents lists available at ScienceDirect

BenchCouncil Transactions on Benchmarks, Standards and Evaluations

journal homepage: www.keaipublishing.com/en/journals/benchcouncil-transactions-onbenchmarks-standards-and-evaluations/

Full Length Article

Medical image fusion based on deep neural network via morphologically processed residuals

Supinder Kaur [a], Parminder Singh [b], Rajinder Vir [b], Arun Singh [b,*], Harpreet Kaur [b]

[a] RBIENT, India
[b] School of Computer Science & Engineering, Lovely Professional University, Phagwara, Punjab, India

ARTICLE INFO

Keywords:
Image fusion
Medical imaging
Image pyramid
Neural network
Residuals

ABSTRACT

Medical image fusion enhances the intrinsic statistical properties of original images by integrating complementary information from multiple imaging modalities, producing a fused representation that supports more accurate diagnosis and effective treatment planning than individual images alone. The principal challenge lies in combining the most informative features without discarding critical clinical details. Although various methods have been explored, it remains difficult to consistently preserve structural and functional features across modalities. To address this, we propose a deep neural network–based framework that incorporates morphologically processed residuals for competent fusion. The network is trained to directly map source images into weight maps thereby overcoming the limitations of traditional activity-level measurements and weight assignment algorithms, and enabling adaptive and reliable weighting of different modalities. The framework further employs image pyramids in a multi-scale design to align with human visual perception, and introduces a local similarity–based adaptive rule for decomposed coefficients to maintain consistency and fine detail preservation. An edge-preserving strategy combining linear low-pass filtering with nonlinear morphological operations is used to emphasize regions of high amplitude and preserve optimally sized structural boundaries. Residuals derived from the linear filter guide the morphological process ensuring significant regions are retained while reducing artifacts. Experimental results demonstrate that the proposed method effectively integrates complementary information from multimodal medical images while mitigating noise, blocking effects, and distortions, leading to fused images with improved clarity and clinical value. This work provides an advanced and reliable fusion approach that contributes substantially to the field of medical image analysis, offering clinicians enhanced visualization tools for decision-making in diagnosis and treatment planning.

1. Introduction

Image fusion technology produces a well-informed single fused image which is highly informative due to combination of multiple source images. In addition to combining image data, this procedure also entails applying one or more algorithms to specifically process the resultant image [1–3]. In medical imaging, fusion is the procedure to combine two or multiple images from several imaging technologies to retain the benefit of complimentary information to generate a more complete image. Due to the emergence of medical imaging technology which includes MRI (magnetic resonance imaging), CT (computed tomography), PET (positron emission tomography), SPECT (single photon emission computed tomography) and several other imaging modalities are widely used in clinical applications and treatment planning. There are many kinds of medical information available from each imaging method. As a useful reference for lesion localization, CT images for instance provide great spatial resolution and good bone imaging. However, soft tissue and fine details of invasive tumours are harder for CT to show. On the other hand, MRI is excellent at soft-tissue imaging which makes it perfect for figuring out how immense a lesion. While their lesser spatial resolution may restrict the ability to diagnose tumours, PET and SPECT can offer useful information about the body's metabolic activities. Depending on a single image type frequently does not yield the best visualization since each imaging modality has inherent limitations resulting from different imaging principles. Medical image fusion can therefore produce more accurate and comprehensive images

Peer review under the responsibility of The International Open Benchmark Council.
* Corresponding author.
E-mail address: arunmandiarun2001@gmail.com (A. Singh).

https://doi.org/10.1016/j.tbench.2025.100237
Received 26 April 2025; Received in revised form 8 September 2025; Accepted 24 September 2025
Available online 26 September 2025

by integrating the complementary information and strengths of several imaging modalities, greatly assisting in diagnosis and treatment planning [4,5].

Recent decades have seen a rise in interest in medical image fusion research, especially in the domain of multiscale transform-based methods. To summarize, the foremost sequential procedures of multiscale transform-based fusion techniques comprise source image decomposition into several coefficients, each of which uses sophisticated operators to encapsulate distinct feature information. In addition, different pixel-level fusion criteria are utilized to combine the matching components and inverse transform is carried out for generating a final fusion result. The initial conventional methods were those that depends on wavelet transforms [6] and Laplacian pyramid transforms [7]. Existing methods acquired diverse image feature information via multiscale transformation, as various frequency components were preserved using identical or simplistic fusion algorithms during the amalgamation of sub-bands. However, a constraint of these approaches is that they analyse only a single image attribute, ignoring others which might significantly reduce the fusion effect [8]. To address this issue, several advanced methods were presented. Tian et al [9] used pixel or regional data to build fusion rules across frequency bands. The fractional wavelet transform approach by Xu et al [10] improves fusion coefficient description. Likewise, Lie et al. presented a sparse representation theory based on minimum spanning trees (MST) [11]. The above-mentioned systems use diverse rules across several frequency bands to get adequate results; however, their rules prevent them from fully integrating detail and contour information. Artifact difficulties are often inevitable due to down sampling and up sampling. To reduce the Gibbs phenomenon, stationary wavelet transform (SWT) image fusion approaches were presented [12], however poor rules hindered fusion of images. Additionally, a shift invariant approaches that include non-subsampled contourlet transform (NSCT) have gained attention. In terms of NSCT approach, Ganasala et al [13] used Laplacian operators and entropy on low and high coefficients. The NSCT based algorithms integrate low and high frequency coefficients via phase congruency, focused contrast, or Laplacian energy restrictions [14]. A popular shift-invariant image decomposition and reconstruction approach is such as nonsubsampled shearlet transform (NSST) uses a shear wave filter. According to Liu et al [15], the gradient factor improves coefficient optimization using the structure tensor and NSST. Singh et al [16] formed a cascaded model using ripplet transform and NSST to enhance direction information. The above methods reveal directional tissue features accurately and comprehensively. The huge difference in parameter efficiency among filters affects the stability of the fusion performance based on parameter selection [17]. This method use pulse coupled neural network (PCNN) which represents a simple subtle vision-based neural network. The global domain PCNN-based approach to image fusion can preserve complex features. However, activating neurons to increase efficiency short of training remains difficult.

Huang et al [18] stimulate each PCNN neuron with the image block's Laplacian energy. An adaptive PCNN method are used in high coefficient region of NSST for increasing PCNN efficiency and fusion quality [19]. Even though there are several PCNN-related fusion approaches, the optimisation of coefficient and threshold characterization are still being explored. Recently, convolutional neural network (CNN) related models are used to solve fusion of medical images problems [20]. The shortages of small sampling data as well as a difficulty for construction of deep neural networks have hampered their consistency. The multiscale domain recommends that CNN sampling or convolution can easily cause fusion phase information loss. Also, a fusion method on the basis of fuzzy radial basis operator neural network is introduced [21], which only activated input neurons using image domain pixel features, which may have limited neural network cognitive capacity and performance. Medical image fusion research emphases on shift-invariant multiscale transformations to create frequency sub-band-specific fusion rules. However, selecting the best fusion rules is difficult. An advanced SR-SCNN blending approach for image fusion [22] contains three phases. After entering complete source images into standard orthogonal matching pursuit, the super-resolution fusion result is generated utilising the max rule to enhance pixel localisation. Besides, each source image receives a special SCNN-based K-SVD dictionary learning approach. A technique improves image information extraction and fusion result sparsity due to its non-linear behaviour. Chao et al [23] develops a novel fusion approach utilising the DSWT and RBFNN. The method first use 2-level decomposition to split two images into seven segments with low- and high frequency sub-bands for DSWT processing. We used the upgraded RBFNN to replace incomplete parts in the same areas of the two images, taking into account the target's gradient and energy. An unsupervised sophisticated image fusion network by Xu et al [24]. This method uses artificial and profound restrictions to boost memory. Saliency and plenty are used to sustain subjective and intuitive qualities in the superficial limitation. The exclusive channels of a pre-trained encoder objectively describe distinguishing information in the deep-level constraint. A multiscale adaptive transformer (MATR) was used for fusion of medical images in unsupervised manner [25]. Instead of regular convolution, adaptive convolution modulates the convolutional layer on the basis of global complementary context. The global semantic extraction was enhanced through an adaptive Transformer, modelling long-range dependencies better. We use a multiscale network design to capture multimodal data at various sizes. Fu et al [26] present a fusion algorithm to fuse medial images. This network comprises fuser, feature extractor along with reconstructor. The feature extractor extracts multiscale features using three MSRPAN blocks. The reconstructor uses three convolutional layers to reconstruct fused features. It also provides the energy ration method for feature fusion. Goyal et al [27] use pixel-based fusion rules to fuse multimodal medical images using cross bilateral and edge-aware filtering. This method calculated final fusion rule weights in a novel way. Both source images are first filtered with a cross bilateral filter (CBF) that considers geometric proximity and neighbouring pixel grey levels. This method prevents edge smoothing by determining the kernel and filtering with one image and vice versa. Subtracting output of CBF from input images yields the detail images. The domain filter extracts smaller scale information from detailed images near large-scale features. A novel image fusion method to solve input image noise and poor contrast [28]. This enhancement algorithm uses CLAHE, BM3D, and the Chameleon Swarm algorithm. An adaptive parameter from the proposed image enhancement approach is utilised for decomposition of image into three augmented layers.

Zhang et al [29] proposes an end-to-end unsupervised learning fusion framework to address these problems. The MMIF implements feature-weighted directed learning for extracting complimentary information from the original images. The feature extraction framework evaluates feature differences at several levels, allowing the feature reconstruction framework for generating interactive weights and directly determine the fusion result. DFENet, a medical image fusion framework by Li et al [30] is a self-supervised blends CNN with vision transformer feature learning. The DFENet's encoder-decoder architecture allows it to train on large natural image datasets without special ground truth fusion images. This network has a feature fuser, encoder as well as decoder. The CNN as well as transformer modules helps for extraction of local and global image information in the encoder. The novel global semantic information aggregation module effectively combines the multiscale characteristics of transformer module by improving image quality and eliminating the need for up-sampling and concatenation. By combining a synthetic focus degree criterion and a specific kernel set, Lepcha et al [31] proposes an image fusion algorithm that performs well in noisy or low-contrast input images. Salient feature extraction initiates with a gaussian curvature filter (GCF) further improve image quality. The dual branch complementary feature injection fusion is proposed by Xie et al [32] using unsupervised CNN models and transformer methods. This method feeds the entire and segmented source images into an adaptive backbone network for capturing local

and global characteristics. As an auxiliary module, the method generates a multi-scale complementary feature extraction framework that emphasizes feature differences at each level for capturing apparent complementary details in source photographs. Song et al [33] present a medical imaging deep learning network model that merges Transformer architecture with an upgraded DenseNet module to overcome the above concerns. The method can be used on natural images. Transformer and dense concatenation reduce feature loss, improving feature extraction and reducing edge blurring.

This study presents a multiscale and pyramid-based approach to produce perceptually better fusion performance based on deep neural network (DNN) with morphological processing of residuals. In particular, each input image is split into Laplacian pyramid and the weight maps constructed from the neural network is split into Gaussian pyramid after morphological processing of residuals. The fusion process is taking place at each decomposition level. Besides, we employ a fusion process on the basis of local similarity for determining the fusion mode for the decomposed coefficients. A weighted-average fusion mode is used when there is a maximum level of similarity between the elements of the input images for preservation of important information. In this case, the weight maps are obtained from the neural network since they are more dependable than a measure based on coefficients. In contrast, a choose-max rule is implemented when there is little image content similarity because it retains the most prominent details from the original images. In this scenario, the outputs of neural network are less dependable, and the absolute values of the decomposed coefficients are utilised to measure pixel activity directly. The network is used to fuse medical images by expanding on these concepts. It is noteworthy that two fundamental methods are utilised in fusion of images i.e., a similarity-based fusion mode determination and the pyramid-based decomposition. In addition, an edge-preserving processing technique is incorporated which selects regions where edges should be retained by combining non-linear techniques with linear lowpass filtering. Based on the morphological processing of liner filter residuals, these regions are selected with the intention of finding important regions with edges that have the right size and high amplitude. Reconstruction operators and area opening are two morphological image processing techniques used to achieve in our method. In order to restore the edges original shape and the identified regions are combined with the results of lowpass filter. In addition, this method permits control over the contrast of they produced image, with four customizable factors influencing the processing result.

The reminder of the paper is organised as follows: In Section 2, we provide a systematic illustration of the proposed fusion method. Section 3 illustrate materials related to datasets, comparative methods, and evaluation measures. In Section 4, an experimental result is presented both visually and quantitatively along with detailed discussion. The final section illustrates the conclusion of the proposed study.

2. Proposed methodology

The network utilised in the proposed fusion approach is shown in Fig. 1. The branches of neural network are inhibited to having the same weights [34]. The convolutional layers as well as max poling layer are composed in each branch. In this network, we eliminate a fully connected layer from the network to construct a much smaller structure with the goal of reducing memory usage and improving computational efficiency. After concatenation, the 512 feature maps are directly connected with the two-dimensional vector. In single accuracy, the smaller mode utilises very less physical memory, which is significantly less than the model used in [34]. In the end, a two-way SoftMax layer receives this two-dimensional vector as source which then generates a probability distribution over two modules. These two modules, first patch 1 and second patch 0, correspondingly first 0 and second patch 1 which represents for two different kinds of normalised weight assessment results. Each probability modules indicates how likely it is that each weight will be assigned. Since the total of the two output probabilities in this case is 1, a possibility of individual module characterizes the weight allocated to each input patch. As described in [34], the network is trained using higher quality image patches and the corresponding blurred fluctuations. Based on the study, the spatial, dimension of the input patch are set to 16×16 during training phase. Random sampling and multiscale gaussian filtering are used in the formation of training. The SoftMax loss function is used as an optimisation goal, which can be reduce utilising the stochastic gradient descent method.

The widely used deep learning structure Caffe is utilised for the training process [35]. Refer [34] for information on network training for the proposed study for proper illustrations. Meanwhile the network includes a fully connected layers whose dimensions are static for both source and the output data since the source of the network needs to have the static size in order to guarantee the source data for the fully connected layers never changes. In case of image fusion, the source image with different sizes can be accommodated by image segmentation into overlapped patches and feeding each pair of patches into the network; however, this procedure significantly increases the number of redundant computations. In order to tackle this issue, the proposed method initially convert fully connected layers into the corresponding conv layer that consists of $8 \times 8 \times 512$ kernel [36]. After conversion, the source images of any size can be processed by the network collectively for generation of dense prediction map. Furthermore, each prediction map which is a two-dimensional vector incorporates the relative clarity details of the source patch pairs at the correlative location. Since that each forecast consists of two dimensions that have been adjusted to a sum of one, the output can reduce to initial weight and the subsequent source. In the end, we allocate the value as the weights for every pixel inside the patch position and computed a mean of the overlapping pixels to produce a weight map that resembles the dimensions of the original images. Fig. 1 demonstrates the schematic representation of proposed fusion algorithm. The algorithm could be described in four stages.

2.1. DNN based weight map generation

Insert two input images A and B into the corresponding neural network branches [37]. Detailed process is described in above section for weight map generation W.

2.2. Morphological processed residuals

This section provides a procedure of the morphological processing of residuals [38,39]. It is related to residual of the gaussian filter:

$$I = W * \mathscr{L} \tag{1}$$

where W stands for a weight map from neural network, * for a convolutional function, and L denotes mask of the gaussian filter (or any other lowpass filter). Besides, the residual of the liner filtering is defined as follows:

$$Res(W) = W - I \tag{2}$$

Operators defined on weight map with positive values are applied to the residual in order to do further processing. As a result, the Res (W), which contains both positive and negative values which can be divided into two parts: positive and negative values.

$$\begin{aligned} I_{res+} &= 0.5(Res(W) + |Res(W)|) \\ I_{res-} &= 0.5(Res(W) - |Res(W)|) \end{aligned} \tag{3}$$

The following apparent relationship is satisfied by the proportion of the residual:

$$Res(W) = I_{res+} + I_{res-} \tag{4}$$

Both fractions of the residual (I_{res+}, I_{res-}) are processed based on the residual's amplitude in order to remove negligible fluctuations while keeping significant ones. The process relies on the morphological

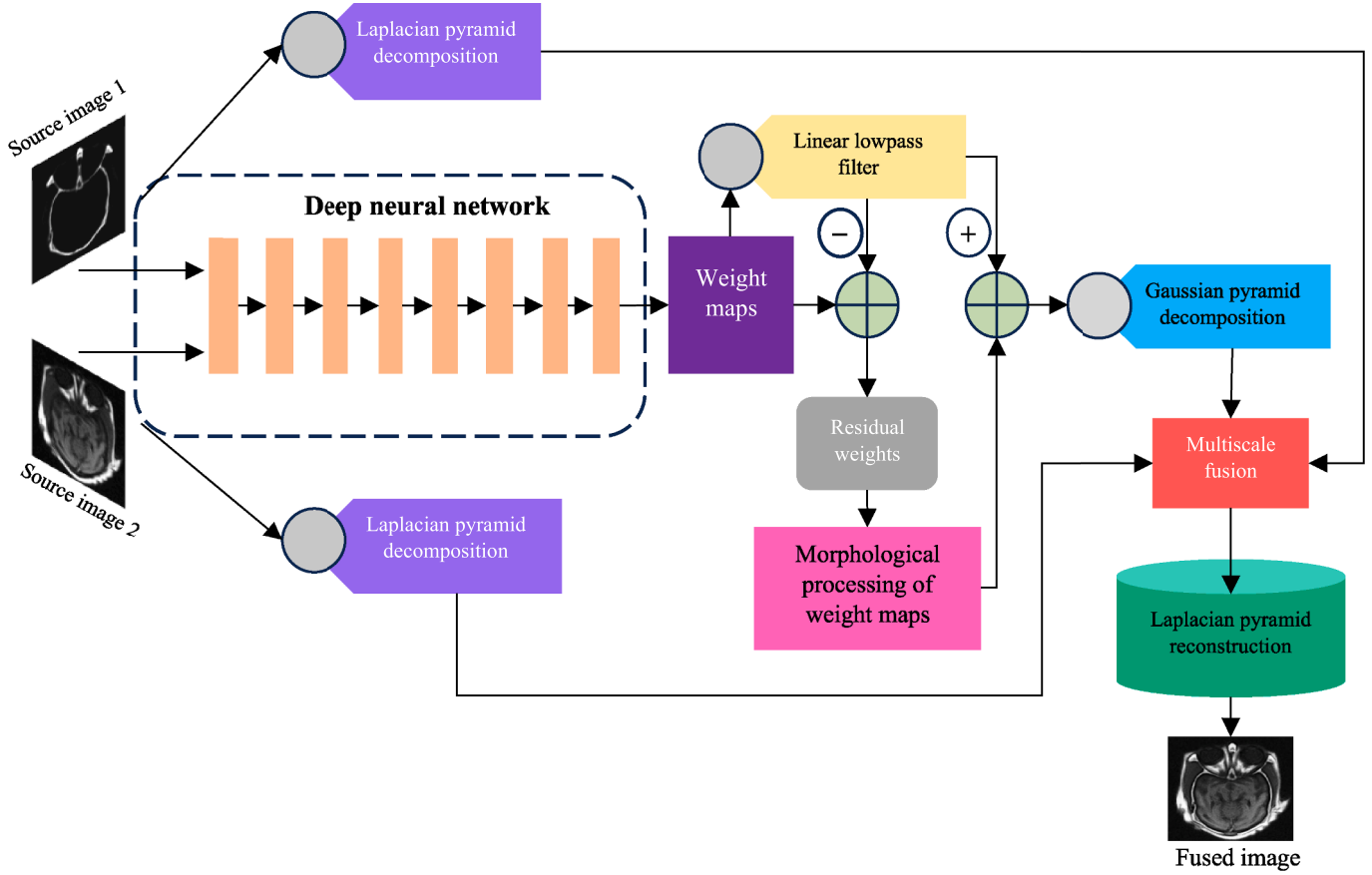


Fig. 1. Methodological flowchart of the proposed method.

function M that maintains the original structures of the residuum while selecting important regions based on the reconstruction. In order to retain relevant borders and blurriness in irrelevant areas, the weight map is finally updated to include large residual regions.

$$I_{out} = I + \mathscr{M}(I_{res+}) - \mathscr{M}(I_{res-}) \tag{5}$$

For both residuals, the operator M is specified as

$$\mathscr{M}(I) = R_I\left(\min\left(I, S_t(I)|_{\{\min\{I\},\ \max\ \{I\}\}}\right)\right) \tag{6}$$

where "|" indicates a mapping operator that generates grey level images by substituting two specified values for the original binary values of 1 s and 0 s, and R_I (A) refers to morphological restoration of the grey level mask I from the marker A. Finally, for two weight maps, '*min*' stands for the pointwise minimal function. Besides, S denotes a mask including significant components, and the concept of M is primarily concerned with contrast preservation (Eq. (6)). A residual amplitude serves as the basis of the substitute of the regions:

$$S_t(I) = \ (I\ \geq t) \tag{7}$$

where S is a selection function that extracts I segments whose amplitude is greater than the specified threshold t. An illustration of image filtering using the previously discussed method was presented in [38]. The test images are filtered using different t values, and the results are shown. Furthermore, a mask is defined as follows: grey denotes regions where both masks equal zero, while white indicates the positive mask $S_t(I_{res(+)})$ and black denotes the negative mask $S_t(I_{res(-)})$. By raising the threshold, fewer components are detected and subsequently restored and used to get the original image data. Ultimately, the quantity of parts displaying their initial clarity is reconstructed. The first function of this method is a mask that is created through thresholding and defines the limits of pertinent elements. A measure of meaningfulness is the amplitude of residuals. The importance of a section of a weight map is defined by more than just amplitude, though. A component of an image with a larger residual amplitude that is unimportant for visual comprehension can be readily imagined. A source image that contains salt and pepper, for example, produces a number of small, high-amplitude image elements, which alters the undesirable details by addition of high amplitude elements. It will be recognized as important but unpleasant places as a result. A further step has been added to our plan to address the previously noted problems. To filter the mask, an area opening filter [40] is utilised. This filter removes all related elements in the image that are less than a given threshold size (i.e. size coefficient). Theus, the expansion of Eq. (7) is described as follows:

$$S_{t,s}(I) = \ (I\ \geq t)\ o\ (s) \tag{8}$$

The area opening which eliminates coefficients smaller than the size represented indicated by s is recognized by the notation $o(s)$. The impact of utilising the size coefficients s and t is illustrated in [38]. It is clear that the number of selected regions reduces as the factors s and t increases. As a result, it allows objects with a small number of pixels to be rejected from the residual even if their amplitude is large. In the end, it makes it easier to maintain these regions in a hazy form in the final weight maps. Addition of (or removal, relies on filter mask coefficients) highpass filtering from the image itself is a conventional process to enhance visual contrasts. This relates to the high pass filtering characteristic that makes it easier to identify local differences in the image pixel values. Alternative technique to obtain high pass filtering outputs is to distinguish between lowpass filter and the image itself. The morphological processing of residuals concerns high frequency components of images which include areas with amplitudes higher than threshold t. Eq. (5) is changed by adding the contrast control coefficient (c) to regulate contrast of the final fused result, which defines in the following formula:

$$W_{out} = I + \ (\mathscr{M}(I_{res+}) - \mathscr{M}(I_{res-})).c. \tag{9}$$

Depending on the value of c, the contrast is either preserved ($c = 1$); enhanced ($c > 1$) or decreased ($0 < c < 1$). Examples indicating how the contrast control coefficient (c) influences the processing results is presented in [38]. Besides, contrast enhancement results are shown without the processing of morphological residuals. The influence of morphological processing on the amount of subtle image details visible in the output is clear in [38]. This processing technique makes it easier to ignore less important information by improving the contrast of relevant weight map regions in images.

2.3. Laplacian and gaussian pyramid decomposition

All the source images are decomposing into a Laplacian pyramid [41]. The pyramids of source images A and B are represented by $S\{A\}^l$ and $S\{B\}^l$, where l is the lth-decomposition level. After morphological processing the residuals of the weight map W, the weight map is further decomposing into a Gaussian pyramid $R\{W\}^l$. For each pyramid, the overall decomposition level is set at the maximum probable values $\lfloor \log_2 \min(H, W) \rfloor$ in which $H \times W$ is the spatial dimensions of the source image and $\lfloor . \rfloor$ indicates the flooring function.

2.4. Fusion of coefficients

A local energy map (i.e. the sum of the squares of coefficients inside a small window) [42] of $S\{A\}^l$ and $S\{B\}^l$, respectively is computed for each decomposition level l [37].

$$\begin{aligned} K_A^l(u,v) &= \sum_m \sum_n S\{A\}^l(u+p,\ v+q)^2, \\ K_B^l(u,v) &= \sum_m \sum_n S\{B\}^l(u+p,\ v+q)^2, \end{aligned} \tag{10}$$

The similarity measure utilised to identify the fusion mode is computed as

$$Z^l(u,v) = \frac{2\ \sum_p \sum_q S\{A\}^l(u+p,\ v+q)\ S\{B\}^l(u+p, v+q),}{K_A^l(u,v) + K_B^l(u,v)} \tag{11}$$

This measure has a range of $[-1,\ 1]$, where values nearer 1 indicate greater similarity. To decide which fusion mode should be used, a threshold t is set. A weighted average fusion mode using weight map W is used if $Z^l(u,v) \geq t$.

$$S\{F\}^l(u,v) = \ R\{W\}^l(u,v)\ .\ S\{A\}^l(u,v) + \ \left(1 - R\{W\}^l(u,v)\right)\ .\ S\{B\}^l(u,v) \tag{12}$$

Eq. (10), which compares the local energy, is utilized to determine the fusion mode if $Z^l(u,v) < \ t$:

$$S\{F\}^l(u,v) = \begin{cases} S\{A\}^l(u,v), \ \text{if}\ K_A^l(u,v)\ \geq\ K_B^l(u,v) \\ S\{B\}^l(u,v), \ \text{if}\ K_A^l(u,v) < \ K_B^l(u,v) \end{cases} \tag{13}$$

The Eq. (14) summarizes the fusion strategy in its entirety. The final fusion result is obtained using Laplacian pyramid reconstruction $S\{F\}^l$.

$$S\{F\}^l(u,v) = \begin{cases} R\{W\}^l(u,v)\ .\ L\{A\}^l(u,v) + \ \left(1 - R\{W\}^l(u,v)\right)\ .\ S\{B\}^l(u,v), \ \text{if}\ Z^l(u,v) \geq t \\ S\{A\}^l(u,v), \ \text{if}\ Z^l(u,v) < t \ \text{and if}\ K_A^l(u,v)\ \geq\ K_B^l(u,v) \\ S\{B\}^l(u,v), \ \text{if}\ Z^l(u,v) < t \ \text{and if}\ K_A^l(u,v) < \ K_B^l(u,v) \end{cases} \tag{14}$$

The adoption of Laplacian and Gaussian pyramid decomposition for multi-scale representation as well as the use of local energy and similarity measures for coefficient analysis are adapted from prior image fusion research. These elements are well-recognized in the literature and have been shown to effectively capture structural details and assess saliency. The novelty of the present work lies in three key aspects. First, we introduce morphological processing of residuals derived from the DNN-generated weight map which allows selective preservation of clinically relevant edges while suppressing noise, with an additional tunable contrast control mechanism for enhancing visibility. Second, we propose a hybrid adaptive fusion rule that dynamically switches between similarity-based weighted fusion and coefficient-based choose-max fusion, thereby integrating the advantages of learned features with robust model-driven decisions. Third, we implement the conversion of fully connected layers to convolutional layers in the training network, enabling efficient generation of dense per-pixel weight maps and avoiding redundant patch-wise processing. Together, these innovations distinguish our approach from conventional pyramid-based or morphology-based fusion methods and constitute the principal contributions of this study.

Algorithm: Medical Image Fusion (DNN + Morphological Residuals)

Input: Two source images A, B
Output: Fused image F

1. Weight Map Generation:
   W = Net.forward(A, B) # DNN produces dense weights
2. Morphological Residual Processing:
   I = conv (W, Gaussian)
   Res = $W - I$
   Res_{pos}, Res_{neg} = split (Res)
   M_{pos} = MorphReconstruct (Res_{pos}, t_{amp}, S_{area})
   M_{neg} = MorphReconstruct (Res_{neg}, t_{amp}, S_{area})
   $W_{ref} = I + c * (\text{M}_{pos} - \text{M}_{neg})$ # contrast control
3. Pyramid Decomposition:
   S_A= LaplacianPyramid(A)
   S_B= LaplacianPyramid(B)
   R_W = GaussianPyramid (W_{ref})
4. Adaptive Fusion (per level l):
   Compute local energy K_A, K_B and similarity Z
   If $Z \geq \tau_{sim}$:
   S_F [l] = R_W [l]* S_A [l] + (1−R_W [l]) * S_B [l]
   Else:
   S_F [l] = max (S_A [l], S_B [l]) by local energy
5. Reconstruction:
   F = LaplacianReconstruct (S_F)
   Return F

3. Materials and data analysis

We used three different types of image dataset pairs (CT-MRI, MRI-SPECT and MRI-PET) to scientifically validate the proposed methodology. Every image is taken from the medical imaging resource The Whole Brain Atlas [43]. Every image is 256 × 256 in size. For each dataset type, we chose 60 image pairings for training and 25 pairs are utilised for testing, i.e. the proposed neural network is trained using 60 pairs of CT-MRI data pairs while using CT-MRI dataset for experiments. We examined 25 data pairs for every type of dataset after training. Nine standard mainstream approaches were used for enhancing the comparison such as enhanced medical image fusion network (EMFusion) [24], discrete stationary wavelet transform and the enhanced radial basis function neural network (DSWT-RBFNN) [23], multiscale adaptive transformer (MATR) [25], feature difference guided network (FDGNet) [29], and dual branch feature enhanced network (DFENet) [30], semantic-preserving fusion (SMFusion) [44], progressive parallel strategy based on deep learning (PPMF-Net) [45], a mamba-based dual--phase model (Mambadfuse) [46]. To improve scientific validation, we assessed each algorithm using five standard quantitative image fusion assessment metrics: Tsallis entropy (TN) [47] that computes the information incorporated in fusion image; mutual information (MI) [4], computes information shared across a fused image and the input images; QAB/F [4], which measures edge similarity based on gradient congruency; spatial frequency (SF) [5], which computes the level of details and textures in the fused image; and the edge preservation index (EPI) [5]. A higher value for each of the above-mentioned criteria denotes a superior fusion performance.

4. Results and discussions

The fused images were generated for every pair of datasets that were considered for fusion as shown in Fig. (2). Figs. (3–9) show the fusion performances of seven dataset pairs using nine different methods. Fig. 3 (c-k) denotes the fusion results generated from the EMFusion, DSWT-RBFNN, MATR, FDGNet, DFENet, Mambadfuse, PPMF-Net SMFusion, and Proposed, respectively. Fig. 3 (a-b) represents two original images considered for fusion. We started by observing at the CT-MRI dataset where Figs. (3–5) show the fusion results of CT-MRI (Dataset 1–3). The CT and MRI (Dataset 1) incorporate of bronchogenic carcinoma. Fig. 3 (c–d) shows the poor fusion effects from the EMFusion and DSWT-RBFNN methods and the brightness of images are greatly distorted as well as the CT contour details are not adequately merged into the fusion performance and the details of soft tissue and bone are not easily noticeable especially with regard to the lesion. The overall fusion outcome remained reasonable even though A MATR-based approach enhanced the information quality representation when comparison to images generated through other two approaches. The features of the lesion are distinctively not visible in MATR method (see Fig. 3(e)). However, as shown in Fig. 3 (f-k), the fusion results of the six methods (i.e. FDGNet, DFENet, Mambadfuse, PPMF-Net, SMFusion and Proposed methods are noticeably better. Significant improvements in structural fidelity and image distortion prevention are made while maintaining the integrity of

A space-occupying sarcoma's pathological changes are represented in the CT-MRI Dataset (2–3). The MRI images represent the edge of the tumour in greater information, whereas the CT images represent the general form of the tumour (see Fig. 4–5). Therefore, it is vital to fuse both the images to accomplished greater detailed images. The proposed algorithm successfully integrates the benefits of CT-MRI images (shown in blue and green circles in Fig. 4 and 5(k) by producing fusion results that are expressively superior than comparative methods. These fused images are illustrious by enhanced image contrast and a complementary broad representation of the edges of the tumour structure. As shown in Figs. (6–7), the fusion results from MRI – SPECT (Dataset 1–2) represents how well the integration of soft tissue and metabolic activity occurs. The intrinsic features and rich feature information of the images are not well represented by the MATR and DFENet based algorithms. In particular, the energy information shows increasing distortion compared to the original SPECT images (Figs. 6 and 7 (c–d)). However, the outputs of the original images generated by the other four algorithms (Mambadfuse, PPMF-Net, SMFusion, and the proposed) presents superior in case of maintaining the energy detail of SPECT images. Besides, when compared of these four approaches, the fusion efficiency of our algorithm clearly outperforms the others by producing sensitive details that are clearer and more comprehensive information features (Red circles in Figs. 9 and 10 (e-*i*) specify the performance of specific details). Figs. (8–9) display the fused images from the MRI-PET Dataset (1–2). Subcortical blood vessel changes are better explained by MRI. A better depiction of metabolism is offered by Positron Emission Tomography (PET). The properties of PECT images are not adequately presented in Figs. 8 and 9 (c–d), which were derived from the MATR and DFENet approaches. There is less contrast as a result of the overall visual brightness being lowered. The fusion consequences are therefore less than ideal. The fusion results of Figs. 8 and 9(h) attained from the proposed algorithm are higher (as shown by red circles) regardless of the energy restoration of the functional images as well as the soft tissues, blood arteries, and

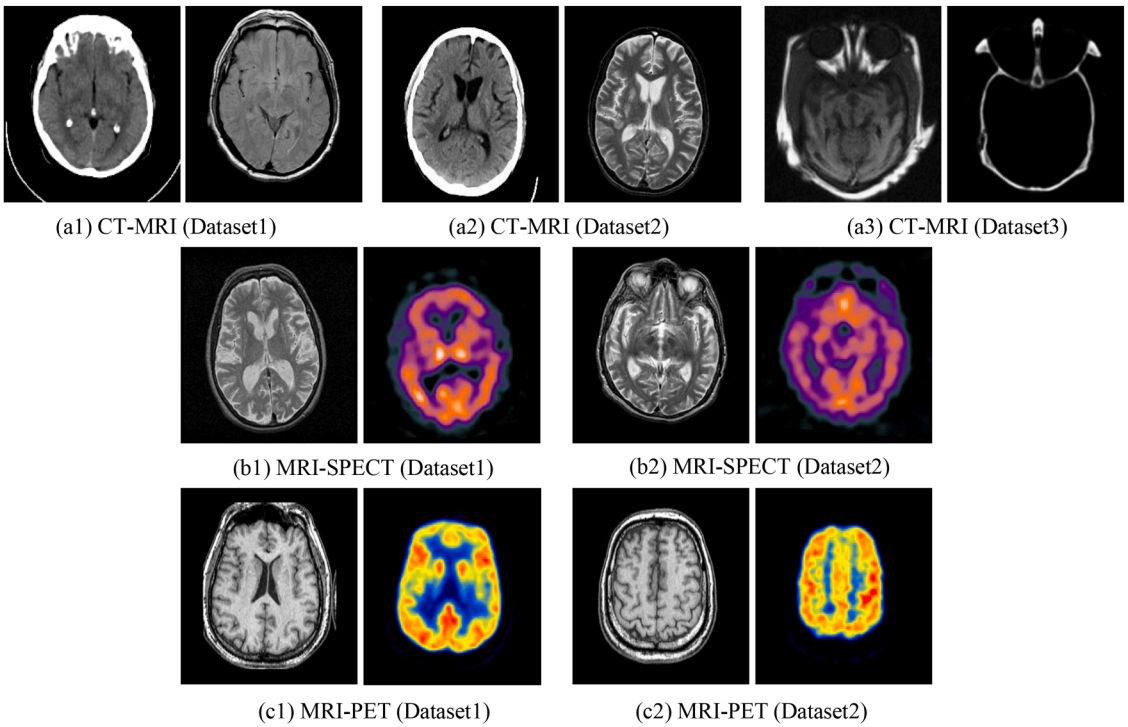


Fig. 2. Experimental pair of source images: (a) CT-MRI; (b) MRI-SPECT and (c) MRI-PET.

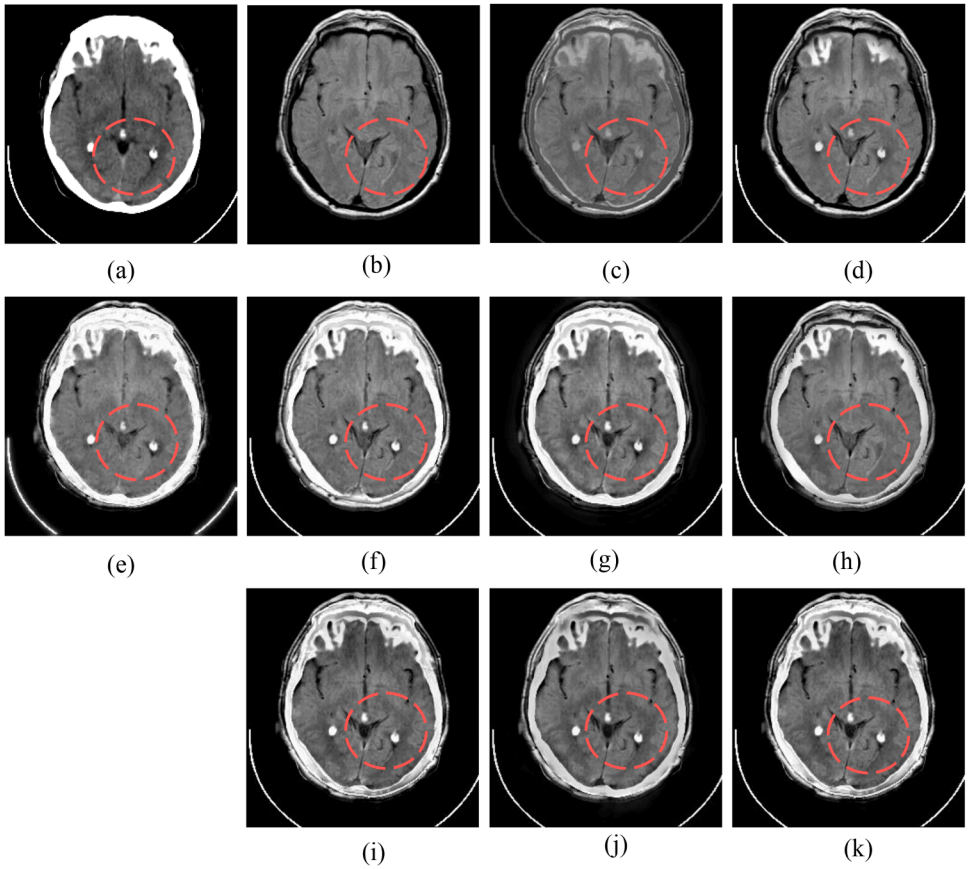


Fig. 3. Fusion performance of CT-MRI Dataset 1 based on nine algorithms: (a) CT; (b) MRI; (c) EMFusion; (d) DSWT-RBFNN; (e) MATR; (f) FDGNet; (g) DFENet; (h) Mambadfuse; (i) PPMF-Net; (j) SMFusion and (k) Proposed, respectively.

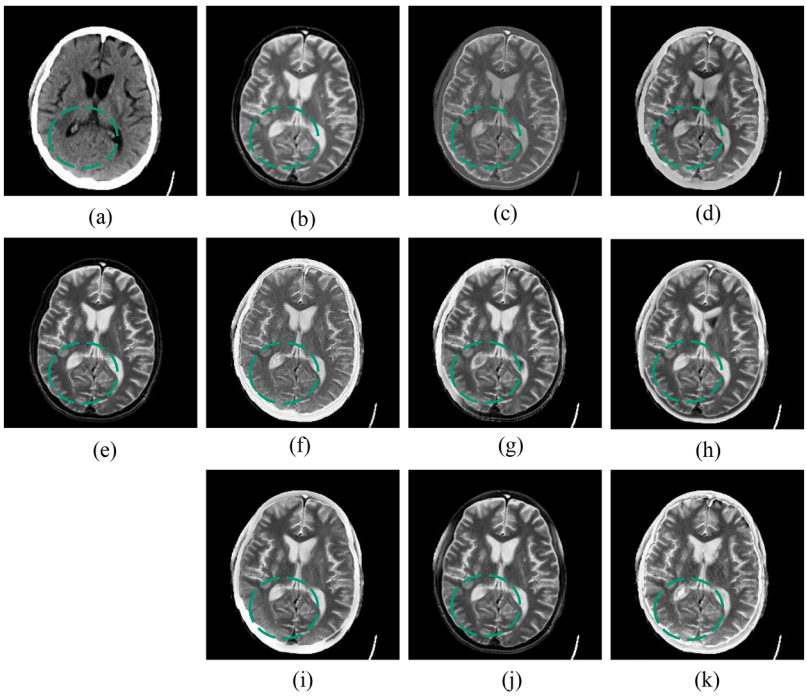


Fig. 4. Fusion performance of CT-MRI Dataset 2 based on nine algorithms: (a) CT; (b) MRI; (c) EMFusion; (d) DSWT-RBFNN; (e) MATR; (f) FDGNet; (g) DFENet; (h) Mambadfuse; (i) PPMF-Net; (j) SMFusion and (k) Proposed, respectively.

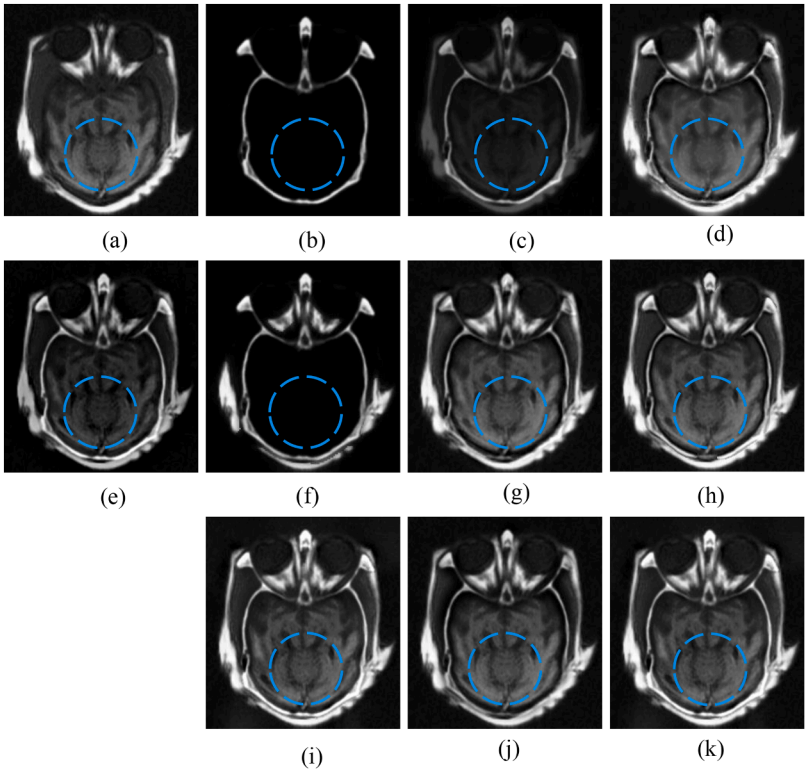


Fig. 5. Fusion performance of CT-MRI Dataset 3 based on nine algorithms: (a) CT; (b) MRI; (c) EMFusion; (d) DSWT-RBFNN; (e) MATR; (f) FDGNet; (g) DFENet; (h) Mambadfuse; (i) PPMF-Net; (j) SMFusion and (k) Proposed, respectively.

other information represented in the MRI images.

The evaluation metrics EPI, MI, SF, TE, and QAB/F were and MRI-PET test data pairs. The mean fusion performance of five metrics across all datasets is shown in Tables (1–3). The objective comprehensive performances obtained from the Mambadfuse, PPMF-Net, SMFusion methods demonstrated good results in terms of visual appraisal, border fusion effect evaluation, and detailed information evaluation. The proposed algorithm outperformed other methods in nearly all of the five metrics such as SF, TE, and QAB/F across all datasets according to a comparison of their performances. The proposed approach also presents

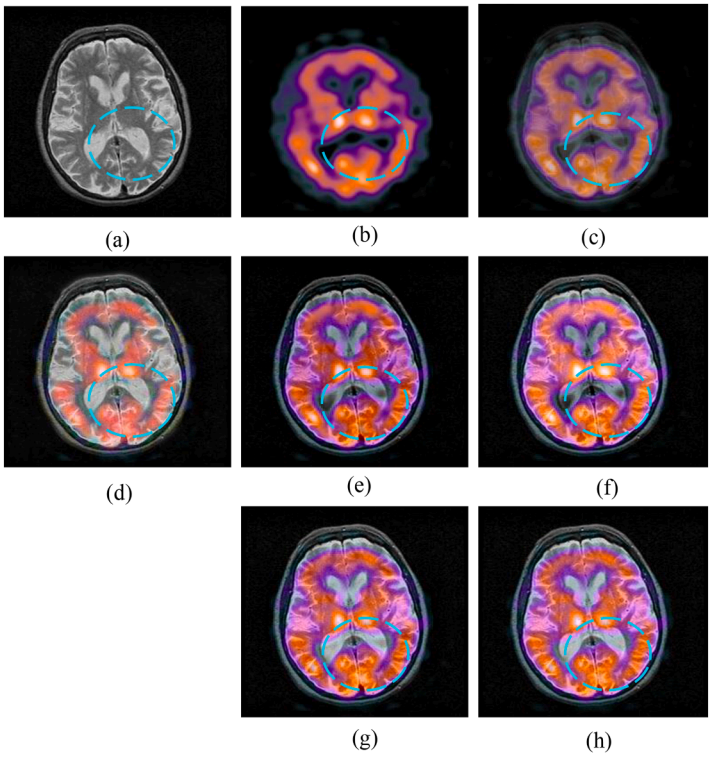


Fig. 6. Fusion performance of MRI-SPECT Dataset 1 based on six algorithms: (a) MRI; (b) SPECT; (c) MATR; (d) DFENet; (e) Mambadfuset; (f) PPMF-Net (g) SMFusion and (h) Proposed, respectively.

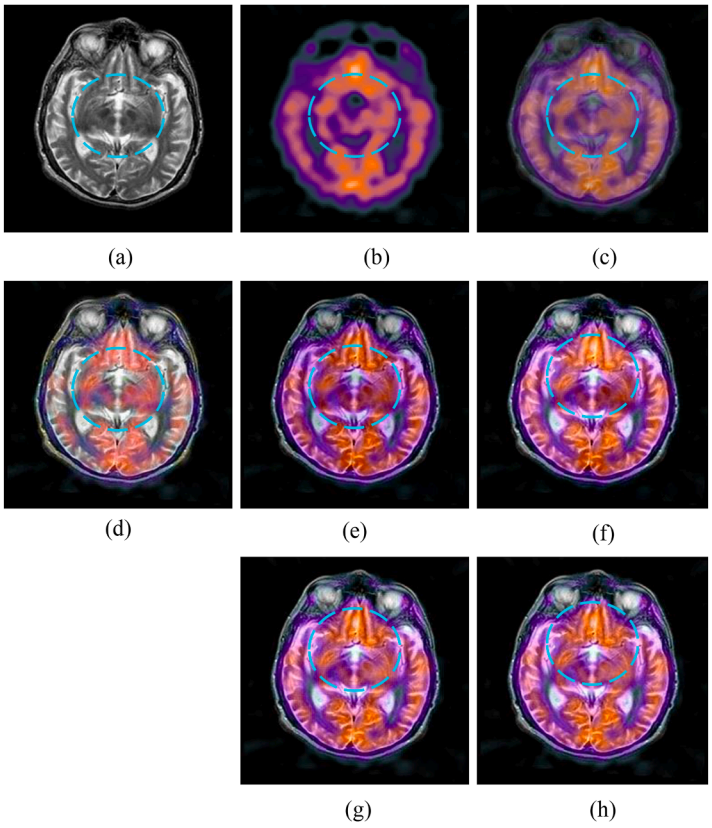


Fig. 7. Fusion performance of MRI and SPECT Dataset 2 based on six algorithms: (a) MRI; (b) SPECT; (c) MATR; (d) DFENet; (e) Mambadfuset; (f) PPMF-Net (g) SMFusion and (h) Proposed, respectively.

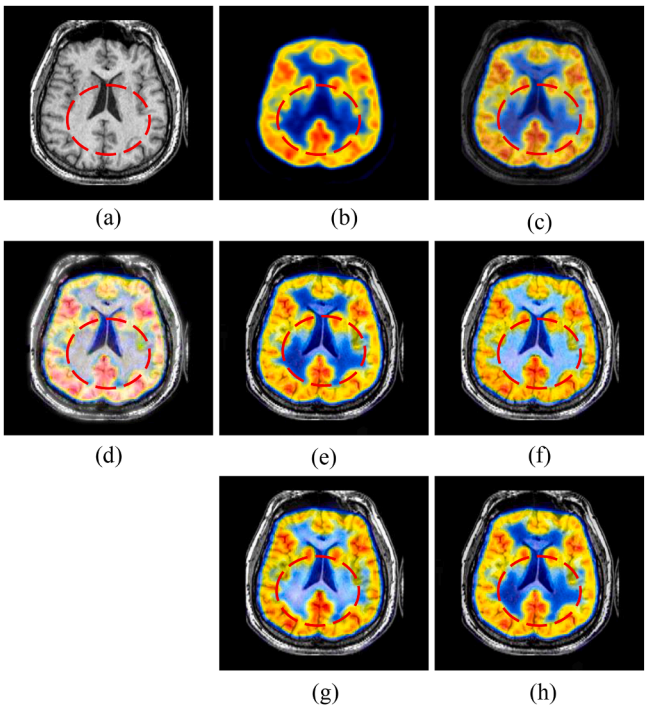


Fig. 8. Fusion performance of MRI-PET Dataset 1 based on six algorithms: (a) MRI; (b) PET; (c) MATR; (d) DFENet; (e) Mambadfuset; (f) PPMF-Net (g) SMFusion and (h) Proposed, respectively.

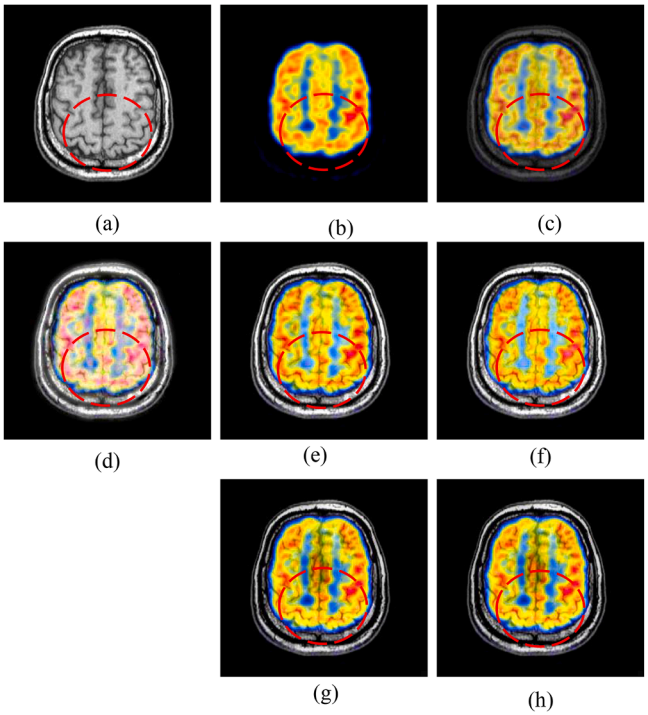


Fig. 9. Fusion performance of MRI and PET Dataset 2 based on six algorithms: (a) MRI; (b) PET; (c) MATR; (d) DFENet; (e) Mambadfuset; (f) PPMF-Net (g) SMFusion and (h) Proposed, respectively.

outstanding performance for EPI and MI for all datasets. The effectiveness of our approach was marginally better to that of PPMF-Net and SMFusion -based methods in all datasets. Subsequently, from a general performance standpoint, the proposed strategy demonstrated superior results in objective analysis which indicates exceptional robustness in fused images. Fig. 10 (a-e) represent an experimental result of different values of Threshold (t) on metrics EPI, MI, QAB/F, SF, TE for CT-MRI images. One can see that when $t = 0.05$, $t = 0.08$, the performance is the better in most scenarios as shown Fig. 10.

The fusion performances of the proposed approach are shown by the final outcomes. The outcomes need to be confirmed by more research that includes a thorough comparison of different imaging types and decomposition levels. The benefits of the DNN-based method occur from its rich deeper architecture and considerable training set. Thus, we maintain the sufficient training data is the key reasons for the improvement of the DNN-based approach. To make the comparison more precise, we trained the DNN based method using the same number of data points as the proposed method. By modifying the neural network design to meet specific experimental needs and choosing an efficient training strategy, we observed that the proposed method permitted the model to exhibit intelligence and self-learning capabilities. The determination of the properties of the input networks is a vital stage in the development of the network design. By combining the pixel and regional block levels of medical image fusion, we were able to achieve a complimentary effect by combining the specific pixel regions and pertinent characteristic properties of the one input image with those of another source image in the same locations. As well as the correlation across pixels and neighbouring pixel values within the regions, the pixel value and the regional energy indicate differences in the properties of pixel characteristics. The selection of these properties from the many subbands gathered by decomposition level is intended to enhance the potentiality of the input layers of network to correctly recognises the signal without causing self-confusion. We used several dimensions in a comparison analysis of 25 pairs of different set of datasets (CT-MRI, MRI-SPECT, MRI-PET) to find the optimal regions. Figures (3–9) shows how a dataset performs visually across different categories. In addition, Tables (1–3) analysed the quantitative performance of the different measures assessed across 25 data pairs shows that our approach produces superior and optimal fusion results. The proposed approach is able to achieved better performance in comparison to recent Mambadfuse, PPMF-Net and SMFusion methods. Medical image fusion the process utilised to complementarily fuse images composed from different medical image technologies. By integrating anatomical as well as functional information, the fusion result presents detailed information from different singe mode imaging modalities without sacrificing any of the energy and information from several images. In the domain of medical image fusion, multiscale transform-based methods are progressively gaining attraction. Also, the efficiency of the fusion procedure is directly impacted by the generation of efficient fusion rules based on the subbands of different coefficients. Despite the simplicity of the fusion rules that are produced by these two approaches, the refinements and features of the original images are not sufficiently recognized and recovered which causes image distortion or degradation. Due to shift variance, the EMFusion method is severely hampered which may result in aliasing and information loss. However, the DSWT-RBFNN based algorithms presented better fusion performance; however, the problem of selecting the best parameters for algorithm activation still exists. There is compromise regarding the utilisation of shift invariant multiscale transforms that is based on processing at the pixel level. Nevertheless, the problems with selection of the appropriate thresholds and parameters or generating suitable fusion rules suffer. We developed fusion rules in this study using neural networks and presented a hybrid approach.

In terms of computational efficiency, the proposed method is designed to minimize overhead by converting the fully connected layers of the network into convolutional layers, which allows dense prediction without redundant patch-wise processing and thereby reduces memory usage and accelerates inference. The additional morphological residual processing step involves only lightweight morphological operators and contributes negligible extra cost. As a result, the inference speed of the proposed approach is competitive with or faster than recent transformer- and dictionary-based fusion methods, while moderately higher

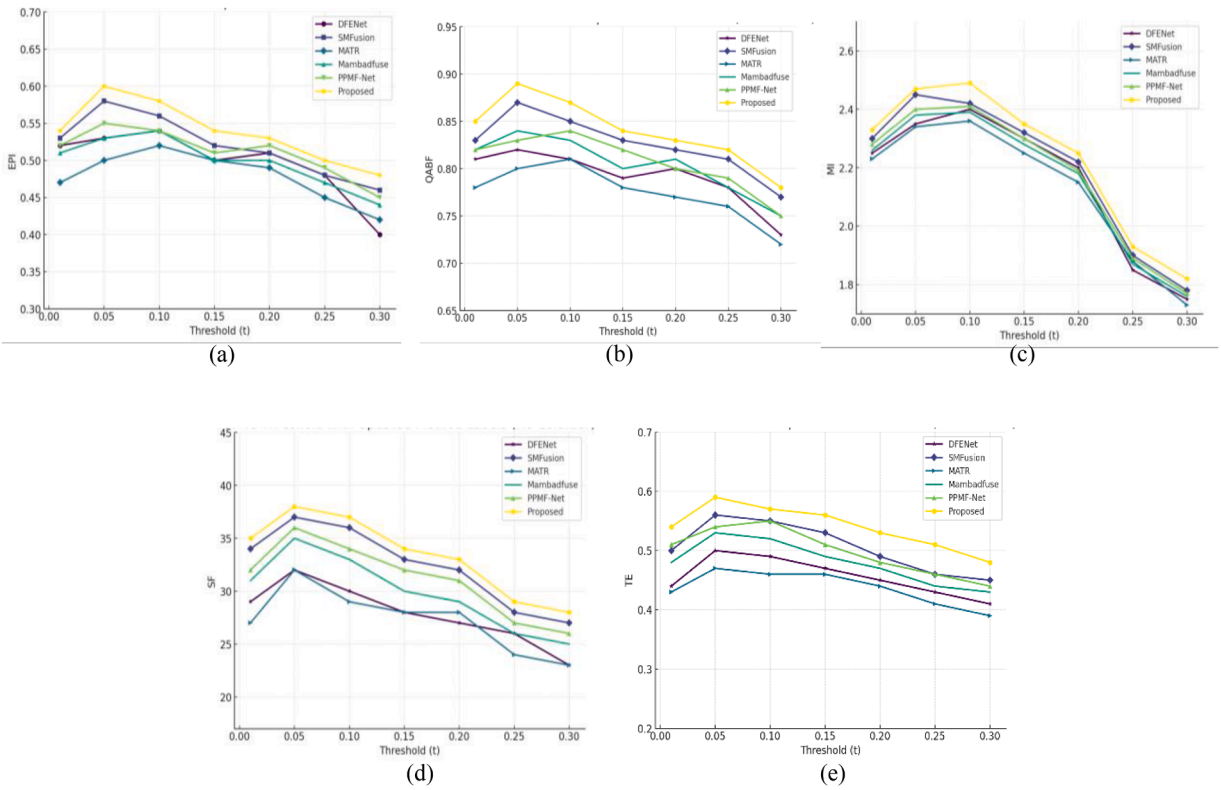


Fig. 10. Experimental performance on different values of Threshold (t) on (a) EPI, (b) QABF, (c) MI, (d) SF, and (e) MRI-PET dataset.

than classical multiscale transforms due to the neural component. Overall, the method achieves a favorable balance between computational cost and fusion performance, ensuring practicality for clinical use.

5. Ablation study

To better understand the contribution of the two core components of the proposed framework such as morphological residual processing (MRP) and the multi-scale pyramid (MSP) design while we carried out an ablation study by selectively removing them from the pipeline and observing the impact on fusion quality. When the morphological residual processing step was excluded, the fused results exhibited reduced sharpness in lesion boundaries and loss of subtle anatomical structures. This that MRP plays a critical role in preserving edges, enhancing local contrast, and suppressing irrelevant noise fluctuations. Its inclusion ensures that diagnostically important regions, such as tumor margins and vascular details, remain well defined in the final fused images. In contrast, removing the multi-scale pyramid design led to a noticeable reduction in the richness of detail and information integration. Single-scale fusion was insufficient to capture both global structural information and fine textural patterns, resulting in fused images with compromised clarity and reduced perceptual quality. The MSP design thus contributes to multi-resolution feature extraction, enabling effective blending of coarse anatomical context with fine localized details. Finally, combining both MRP and MSP provided the most balanced results, delivering fused images with high structural fidelity, strong edge preservation, and comprehensive detail integration. This demonstrates that the two components are complementary and mutually reinforcing, and together they form the backbone of the robustness and clinical relevance of the proposed method.

A limitation of this study is the relatively small training set (60 pairs per modality), which could restrict feature diversity and raise concerns about generalization. To address this, we adopted patch-based training with overlapping 16 × 16 patches and integrated morphological residual processing, reducing dependence on purely data-driven learning.

Table 1
Quantitative evaluation of CT-MRI fusion images.

m	EPI	MI	$Q^{AB/F}$	SF	TE
EMFusion	0.3562	1.4425	0.5672	26.8715	0.3828
DSWT-RBFNN	0.3892	1.5626	0.6344	28.8162	0.4081
MATR	0.4252	1.6751	0.6726	32.8816	0.4312
FDGNet	0.4362	1.7718	0.7125	31.8373	0.4971
DFENet	0.4692	1.8826	0.7552	33.0081	0.4816
Mambadfuse	0.5352	1.9727	0.8077	34.8811	0.5244
PPMF-Net	0.5072	1.9923	0.8252	36.8817	0.5517
SMFusion	0.4987	2.0028	0.8352	35.9817	0.5431
Proposed	0.5862	2.4596	0.8445	37.8622	0.5775

Table 2
Quantitative evaluation of MRI-SPECT fusion images.

	EPI	MI	$Q^{AB/F}$	SF	TE
MATR	0.4627	2.5381	0.7625	17.9928	0.2344
DFENet	0.5177	2.6715	0.7982	19.9172	0.2554
Mambadfuse	0.4926	3.0018	0.8026	21.4413	0.2617
PPMF-Net	0.5673	3.2891	0.8055	20.1872	0.2803
SMFusion	0.5424	3.4324	0.8141	21.9272	0.2781
Proposed	0.5813	3.6772	0.8261	23.8229	0.2897

Table 3
Quantitative evaluation of MRI-PET fusion images.

	EPI	MI	$Q^{AB/F}$	SF	TE
MATR	0.4625	1.9972	72.9937	24.9198	0.3927
DFENet	0.4972	2.1888	74.8838	25.8827	0.4366
Mambadfuse	0.5193	2.3233	77.8272	27.1993	0.4628
PPMF-Net	0.5623	2.5728	80.9727	28.1433	0.4902
SMFusion	0.5565	2.5552	81.8827	28.3837	0.5321
Proposed	0.5896	2.6954	83.8272	30.5286	0.5451

These strategies, along with consistent improvements across CT-MRI, MRI-SPECT, and MRI-PET datasets (Tables 1–3), indicate that the model generalizes well despite limited data. Nonetheless, future work will expand training using larger public datasets, apply augmentation and transfer learning, and validate on multi-center clinical images to further strengthen robustness and applicability.

6. Conclusion

This paper introduces a robust medical image fusion framework that leverages a deep neural network in combination with morphological processing of residuals. The core of the approach lies in the generation of adaptive weight maps through a deep neural network, which effectively captures and integrates pixel-level activity information from the source images. To align the fusion process with human visual perception, image pyramids are employed to achieve multiscale representation, ensuring consistent integration of both fine and coarse details. The fusion mode is further refined using a local similarity–based strategy that adaptively adjusts for decomposed image components. A key element of the methodology is the incorporation of edge-preserving processing, designed to maintain critical structural boundaries. This is accomplished by combining nonlinear algorithms with linear low-pass filtering to identify and preserve regions with high amplitude and optimally scaled edges. The morphologically processed residuals derived from linear filter outputs serve as a reliable basis for selecting significant regions. Morphological operations such as reconstruction and area opening are employed to retain the original shape of edges, while blending the low-pass filter results with the selected regions ensures accurate structural preservation. Moreover, the method provides flexibility for contrast adjustment in the fused image, enhancing its diagnostic utility. Comprehensive experiments on diverse medical datasets demonstrate that the proposed framework outperforms several state-of-the-art techniques, delivering fused results with improved information richness, structural fidelity, and preservation of fine details from the source modalities.

Data availability

The dataset used in this study is publicly available from the Harvard Medical School MIF Database and can be accessed at: http://www.med.harvard.edu/AANLIB/home.html.

Funding

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

CRediT authorship contribution statement

Supinder Kaur: Conceptualization, Data curation, Formal analysis, Funding acquisition. **Parminder Singh:** Formal analysis, Investigation. **Rajinder Vir:** Project administration, Resources, Software. **Arun Singh:** Writing – review & editing, Writing – original draft, Methodology. **Harpreet Kaur:** Supervision, Validation.

Declaration of competing interest

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.

Acknowledgments

We are thankful to the reviewers and editors for their valuable suggestions. We appreciate their insightful comments which helped us in improving the manuscript significantly.

References

[1] M.A. Azam, et al., A review on multimodal medical image fusion: compendious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics, Comput. Biol. Med. 144 (May 2022) 105253, https://doi.org/10.1016/J.COMPBIOMED.2022.105253.
[2] S. Ullah Khan, M. Ahmad Khan, M. Azhar, F. Khan, Y. Lee, M. Javed, Multimodal medical image fusion towards future research: a review, J. King Saud Univ. - Comput. Inf. Sci. 35 (8) (Sep. 2023) 101733, https://doi.org/10.1016/J.JKSUCI.2023.101733.
[3] S. Basu, S. Singhal, D. Singh, A systematic literature review on multimodal medical image fusion, Multimed. Tools. Appl. 83 (6) (Feb. 2024) 15845–15913, https://doi.org/10.1007/S11042-023-15913-W/TABLES/9.
[4] B. Goyal, D. Chyophel Lepcha, A. Dogra, V. Bhateja, A. Lay-Ekuakille, Measurement and analysis of multi-modal image fusion metrics based on structure awareness using domain transform filtering, Measurement 182 (Sep. 2021) 109663, https://doi.org/10.1016/J.MEASUREMENT.2021.109663.
[5] D.C. Lepcha, et al., Multimodal medical image fusion based on pixel significance using anisotropic diffusion and cross bilateral filter, Hum.-centric Comput. Inf. Sci. 12 (2022), https://doi.org/10.22967/HCIS.2022.12.015.
[6] G. Pajares, J.M. de la Cruz, A wavelet-based image fusion tutorial, Pattern. Recognit. 37 (9) (Sep. 2004) 1855–1872, https://doi.org/10.1016/J.PATCOG.2004.03.010.
[7] P.J. BURT, E.H. ADELSON, The Laplacian Pyramid as a Compact Image Code. Readings in Computer Vision, Jan. 1987, pp. 671–679, https://doi.org/10.1016/B978-0-08-051581-6.50065-9.
[8] Y. Yang, D.S. Park, S. Huang, N. Rao, Medical image fusion via an effective wavelet-based approach, EURASIP. J. Adv. Signal. Process. 2010 (1) (Apr. 2010) 1–13, https://doi.org/10.1155/2010/579341. *2010 2010:1*.
[9] H. Tian, Y.N. Fu, P.G. Wang, Image fusion algorithm based on regional variance and multi-wavelet bases, in: Proceedings of the 2010 2nd International Conference on Future Computer and Communication, ICFCC 2010 2, 2010, https://doi.org/10.1109/ICFCC.2010.5497628.
[10] X. Xu, Y. Wang, S. Chen, Medical image fusion using discrete fractional wavelet transform, Biomed. Signal. Process. Control 27 (May 2016) 103–111, https://doi.org/10.1016/J.BSPC.2016.02.008.
[11] Y. Liu, S. Liu, Z. Wang, A general framework for image fusion based on multi-scale transform and sparse representation, Inf. Fusion 24 (Jul. 2015) 147–164, https://doi.org/10.1016/J.INFFUS.2014.09.004.
[12] P. Ganasala, A.D. Prasad, Medical image fusion based on laws of texture energy measures in stationary wavelet transform domain, Int. J. ImAging Syst. Technol. 30 (3) (Sep. 2020) 544–557, https://doi.org/10.1002/IMA.22393.
[13] P. Ganasala, V. Kumar, CT and MR image fusion scheme in nonsubsampled contourlet transform domain, J. Digit. Imaging 27 (3) (Jan. 2014) 407–418, https://doi.org/10.1007/S10278-013-9664-X/METRICS.
[14] Z. Zhu, M. Zheng, G. Qi, D. Wang, Y. Xiang, A phase congruency and local LAPLACIAN energy based multi-modality medical image fusion method in NSCT domain, IEEe Access. 7 (2019) 20811–20824, https://doi.org/10.1109/ACCESS.2019.2898111.
[15] X. Liu, W. Mei, H. Du, Structure tensor and nonsubsampled shearlet transform based algorithm for CT and MRI image fusion, Neurocomputing. 235 (Apr. 2017) 131–139, https://doi.org/10.1016/J.NEUCOM.2017.01.006.
[16] S. Singh, R.S. Anand, D. Gupta, CT and MR image information fusion scheme using a cascaded framework in ripplet and NSST domain, IET. Image Process. 12 (5) (May 2018) 696–707, https://doi.org/10.1049/IET-IPR.2017.0214.
[17] D. Gai, X. Shen, H. Cheng, H. Chen, Medical image fusion via PCNN Based on edge preservation and improved sparse representation in NSST domain, IEEe Access. 7 (2019) 85413–85429, https://doi.org/10.1109/ACCESS.2019.2925424.
[18] W. Huang, Z. Jing, Multi-focus image fusion using pulse coupled neural network, Pattern. Recognit. Lett. 28 (9) (Jul. 2007) 1123–1132, https://doi.org/10.1016/J.PATREC.2007.01.013.
[19] C. Panigrahy, A. Seal, N.K. Mahato, MRI and SPECT image fusion using a weighted parameter adaptive dual channel PCNN, IEEe Signal. Process. Lett. 27 (2020) 690–694, https://doi.org/10.1109/LSP.2020.2989054.
[20] K. Wang, M. Zheng, H. Wei, G. Qi, Y. Li, Multi-modality medical image fusion using convolutional neural network and contrast pyramid, Sensors 20 (8) (Apr. 2020) 2169, https://doi.org/10.3390/S20082169, 2020, Vol. 20, Page 2169.
[21] Z. Chao, D. Kim, H.J. Kim, Multi-modality image fusion based on enhanced fuzzy radial basis function neural networks, Physica Medica 48 (Apr. 2018) 11–20, https://doi.org/10.1016/J.EJMP.2018.03.008.
[22] A. Sabeeh Yousif, Z. Omar, U.Ullah Sheikh, An improved approach for medical image fusion using sparse representation and Siamese convolutional neural network, Biomed. Signal. Process. Control 72 (Feb. 2022) 103357, https://doi.org/10.1016/J.BSPC.2021.103357.
[23] Z. Chao, X. Duan, S. Jia, X. Guo, H. Liu, F. Jia, Medical image fusion via discrete stationary wavelet transform and an enhanced radial basis function neural network, Appl. Soft. Comput. 118 (Mar. 2022) 108542, https://doi.org/10.1016/J.ASOC.2022.108542.
[24] H. Xu, J. Ma, EMFusion: an unsupervised enhanced medical image fusion network, Inf. Fusion 76 (Dec. 2021) 177–186, https://doi.org/10.1016/J.INFFUS.2021.06.001.
[25] W. Tang, F. He, Y. Liu, Y. Duan, MATR: multimodal medical image fusion via multiscale adaptive transformer, IEEE Trans. Image Process. 31 (2022) 5134–5149, https://doi.org/10.1109/TIP.2022.3193288.

[26] J. Fu, W. Li, J. Du, Y. Huang, A multiscale residual pyramid attention network for medical image fusion, Biomed. Signal. Process. Control 66 (Apr. 2021) 102488, https://doi.org/10.1016/J.BSPC.2021.102488.
[27] B. Goyal, et al., Multi-modality image fusion for medical assistive technology management based on hybrid domain filtering, Expert. Syst. Appl. 209 (Dec. 2022) 118283, https://doi.org/10.1016/J.ESWA.2022.118283.
[28] P.H. Dinh, Medical image fusion based on enhanced three-layer image decomposition and Chameleon swarm algorithm, Biomed. Signal. Process. Control 84 (Jul. 2023) 104740, https://doi.org/10.1016/J.BSPC.2023.104740.
[29] G. Zhang, R. Nie, J. Cao, L. Chen, Y. Zhu, FDGNet: a pair feature difference guided network for multimodal medical image fusion, Biomed. Signal. Process. Control 81 (Mar. 2023) 104545, https://doi.org/10.1016/J.BSPC.2022.104545.
[30] W. Li, Y. Zhang, G. Wang, Y. Huang, R. Li, DFENet: a dual-branch feature enhanced network integrating transformers and convolutional feature learning for multimodal medical image fusion, Biomed. Signal. Process. Control 80 (Feb. 2023) 104402, https://doi.org/10.1016/J.BSPC.2022.104402.
[31] D.C. Lepcha, B. Goyal, A. Dogra, A. Alkhayyat, S.K. Shah, V. Kukreja, A robust medical image fusion based on synthetic focusing degree criterion and special kernel set for clinical diagnosis, J. Comput. Sci. 20 (4) (Feb. 2024) 389–399, https://doi.org/10.3844/JCSSP.2024.389.399.
[32] Y. Xie, L. Yu, C. Ding, CFIFUSION: dual-branch complementary feature injection network for medical image fusion, Int. J. ImAging Syst. Technol. 34 (4) (Jul. 2024) e23144, https://doi.org/10.1002/IMA.23144.
[33] Y. Song, et al., DESTRANS: a medical image fusion method based on transformer and improved DENSENET, Comput. Biol. Med. 174 (May 2024) 108463, https://doi.org/10.1016/J.COMPBIOMED.2024.108463.
[34] Y. Liu, X. Chen, H. Peng, Z. Wang, Multi-focus image fusion with a deep convolutional neural network, Inf. Fusion 36 (Jul. 2017) 191–207, https://doi.org/10.1016/J.INFFUS.2016.12.001.
[35] Y. Jia, et al., Caffe: convolutional architecture for fast feature embedding, in: Proceedings of the MM 2014 - Proceedings of the 2014 ACM Conference on Multimedia, Nov. 2014, pp. 675–678, https://doi.org/10.1145/2647868.2654889.
[36] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, Y. LeCun, OverFeat: integrated recognition, localization and detection using convolutional networks, in: Proceedings of the International Conference on Learning Representations, 2013.
[37] Y. Liu, X. Chen, J. Cheng, H. Peng, A medical image fusion method based on convolutional neural networks, in: Proceedings of the 20th International Conference on Information Fusion, Fusion 2017 - Proceedings, Aug. 2017, https://doi.org/10.23919/ICIF.2017.8009769.
[38] M. Iwanowski, "Edge-Aware Color Image Manipulation by Combination of Low-Pass Linear Filter and Morphological Processing of Its Residuals," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 12334 LNCS, pp. 59–71, 2020, doi: 10.1007/978-3-030-59006-2_6.
[39] D.C. Lepcha, B. Goyal, A. Dogra, S.H. Wang, J.S. Chohan, Medical image enhancement strategy based on morphologically processing of residuals using a special kernel, Expert. Syst. (2022) e13207, https://doi.org/10.1111/EXSY.13207.
[40] P. Soille, Morphological image analysis, Morphol. Image Anal. (2004), https://doi.org/10.1007/978-3-662-05088-0.
[41] W. Wang and F. Chang, "A multi-focus image fusion method based on laplacian pyramid," 2011, doi: 10.4304/jcp.6.12.2559-2566.
[42] P.J. Burt, R.J. Kolczynski, Enhanced image capture through fusion, in: Proceedings of the 1993 IEEE 4th International Conference on Computer Vision, 1993, pp. 173–182, https://doi.org/10.1109/ICCV.1993.378222.
[43] "The Whole Brain Atlas." Accessed: Oct. 12, 2024. [Online]. Available: https://www.med.harvard.edu/aanlib/home.html.
[44] Haozhe Xiang, Han Zhang, Yu Cheng, Xiongwen Quan, and Wanwan Huang. "SMFUSION: semantic-preserving fusion of multimodal medical images for enhanced clinical diagnosis." Arxiv Preprint *arXiv:2505.12251* (2025).
[45] Peng Peng, Yaohua Luo, Multimodal medical image fusion using a progressive parallel strategy based on deep learning, Electronics. (Basel) 14 (11) (2025) 2266.
[46] Zhe Li, Haiwei Pan, Kejia Zhang, Yuhua Wang, and Fengming Yu. "Mambadfuse: a mamba-based dual-phase model for multi-modality image fusion." Arxiv Preprint *arXiv:2404.08406* (2024).
[47] A. Sholehkerdar, J. Tavakoli, Z. Liu, Theoretical analysis of TSALLIS entropy-based quality measure for weighted averaging image fusion, Inf. Fusion 58 (Jun. 2020) 69–81, https://doi.org/10.1016/J.INFFUS.2019.12.010.

BenchCouncil Transactions on Benchmarks, Standards and Evaluations 5 (2025) 100245

Contents lists available at ScienceDirect

BenchCouncil Transactions on Benchmarks, Standards and Evaluations

journal homepage: www.keaipublishing.com/en/journals/benchcouncil-transactions-on-benchmarks-standards-and-evaluations/

Research article

Evaluatology-driven artificial intelligence

Guoxin Kang [b], Wanling Gao [a,b,*], Jianfeng Zhan [a,b,c]

[a] The International Open Benchmark Council, China
[b] ICT, Chinese Academy of Sciences, Beijing, China
[c] University of Chinese Academy of Sciences, Beijing, China

ARTICLE INFO

Keywords:
Evaluatology
Artificial intelligence

ABSTRACT

The prevailing data-driven paradigm in AI has largely neglected the generative nature of data. All data, whether observational or experimental, are produced under specific conditions, yet current approaches treat them as context-free artifacts. This neglect results in uneven data quality, limited interpretability, and fragility when models face novel scenarios. Evaluatology reframes evaluation as the process of inferring the influence of an evaluated object on the affected factors and attributing the evaluation outcome to specific ones. Among these factors, a minimal set of indispensable elements determines how changes in conditions propagate to outcomes. This essential set constitutes the evaluation conditions. Together, the evaluated object and its evaluation conditions form a self-contained evaluation system — a structured unit that anchors evaluation to its essential context.

We propose an evaluatology-based paradigm, which spans the entire AI lifecycle — from data generation to training and evaluation. Within each self-contained evaluation system, data are generated and distilled into their invariant informational structures. These distilled forms are abstracted into reusable causal-chain schemas, which can be instantiated as training examples. By explicitly situating every learning instance within such condition-aware systems, evaluation is transformed from a passive, post-hoc procedure into an active driver of model development. This evaluation-based paradigm enables the construction of causal training data that are interpretable, traceable, and reusable, while reducing reliance on large-scale, unstructured datasets. This paves the way toward scalable, transparent, and epistemically grounded AI.

1. Introduction

In artificial intelligence, the forms of data are endlessly diverse — spanning text, images, audio, and structured records — yet their underlying essence remains invariant [1]. This *nature of data* represents the stable informational structure that persists beneath superficial variations. However, the prevailing data-driven paradigm has largely overlooked this essence, instead pursuing a brute-force strategy of enumerating variations to cover possible conditions. Such an approach, even when exhaustive, remains brittle in the face of unseen scenarios, as it fails to address the causal factors that truly govern generalization [2].

However, The continued viability of data-driven AI is now being challenged by deep scarcity limitations. As models grow larger and more data-hungry, the supply of high-quality, human-authored training data has become a critical bottleneck [3]. Recent projections by Epoch AI warn that the stock of usable Internet-scale text data may be depleted by 2028 [4]. This looming *data ceiling* threatens not only the scalability of current systems but also undermines their epistemic reliability and generalization capabilities [5]. In response, many have turned to generative models to produce synthetic data in an attempt to overcome this limitation [6–10]. However, such data often deviates significantly from real-world distributions, introducing distributional shifts that compromise its effectiveness as training material. Coupled with growing concerns over bias, opacity, hallucination, and uncontrollable behaviors — problems exacerbated by the black-box nature of large models and their dependence on opaque training corpora — the limitations of the current paradigm are becoming increasingly apparent [11].

We argue that a fundamental shift is needed: from scaling data volume to capturing and leveraging the nature of data itself — its invariant informational essence beneath diverse forms. To this end, we propose an *evaluation-based paradigm*, in which learning is guided and constrained by well-defined evaluation conditions rather than by uncontrolled data accumulation [12,13]. Drawing from the emerging discipline of Evaluatology, we reimagine the AI lifecycle such that evaluation is not a terminal procedure but a core methodological principle, shaping data generation and governing training dynamics. This

* Corresponding author at: ICT, Chinese Academy of Sciences, Beijing, China.
E-mail address: gaowanling@ict.ac.cn (W. Gao).

https://doi.org/10.1016/j.tbench.2025.100245
Received 28 August 2025; Received in revised form 26 September 2025; Accepted 30 September 2025
Available online 15 October 2025

reconceptualization is both technical and epistemological, redefining how intelligence should be constructed, assessed, and trusted in the post-data-scaling era.

At the heart of this paradigm are Self-contained Evaluation Systems (SESs): self-contained units that produce data — both observational and experimental — under explicit evaluation conditions. Within each SES, data are produced under explicit evaluation conditions and then distilled into their essential informational structures, from which causal-chain schemas are abstracted and later instantiated into concrete training examples. This process guarantees that the resulting data are interpretable, causally grounded, and reusable across tasks.

When applied to large language model development, SESs enable the construction of causal training data — for instance, reasoning traces, grounding steps, or labeled decision points — that make the provenance of every learning instance explicit. In contrast to opaque, data-hungry pipelines, our approach yields a white-box training system in which condition changes can be precisely traced to output variations. This ensures interpretability, attribution, and transparency while significantly reducing dependence on large-scale unstructured datasets.

2. Related work and motivation

2.1. Related work

2.1.1. The existing AI paradigm

As shown in Fig. 1(a), early developments in data-driven artificial intelligence relied heavily on manually labeled data, which enabled supervised learning for tasks such as classification and regression. While effective in domains like computer vision [14] and natural language processing [15], labeled data incurs high annotation costs, suffers from limited scalability, and is often prone to label noise [16], limiting its applicability in large-scale or open-ended scenarios.

To overcome these limitations, researchers have increasingly turned to unlabeled data, leveraging unsupervised and self-supervised learning techniques to extract patterns and representations without explicit human annotations. Classical methods include clustering [17], dimensionality reduction [18], and association mining. Recent advances in self-supervised learning — such as masked language modeling [19], contrastive learning [20], and masked image modeling [21] — have demonstrated strong performance across modalities and become foundational for large-scale pretraining.

From the perspective of learning paradigms, supervised and unsupervised learning are often sufficient for solving simple tasks. However, as task complexity increases — e.g., involving dynamic environments, delayed feedback, or alignment with human intent — these paradigms are frequently integrated with reinforcement learning to enable adaptive, multi-stage training pipelines. This hybrid approach has become central to the development of general-purpose systems such as large language models, where unsupervised learning supports knowledge acquisition, supervised learning guides task-specific behavior, and reinforcement learning fine-tunes alignment with user preferences.

In this work, we focus on static data sources. While, reinforcement learning, which depends on interaction-driven data collection and reward-based optimization, follows a fundamentally different data paradigm and is therefore excluded from our data-centric analysis [22–24].

2.1.2. The basic concepts, theories, and methodologies in evaluatology

Evaluatology [12,13] conceptualizes evaluation as the process of inferring the effect of an evaluated object on the affected objects. An object naturally exerts influence on multiple others, which can be categorized as directly or indirectly affected. Based on this principle, Evaluatology stresses the need to specify evaluation conditions (ECs) and to identify the set of affected objects. Together, the evaluated object, the affected objects, and the evaluation conditions form a Self-contained Evaluation System (SES). For instance, in evaluating CPU performance, the CPU serves as the evaluated object, while stakeholders' concerns — such as running time — require consideration of affected objects including the dataset, algorithm implementation, programming framework, operating system, compiler, processor, memory, etc [25]. Each possible configuration of these affected objects constitutes a point in a vast evaluation condition space, where variations in conditions naturally lead to variations in running time results. This structured perspective highlights that evaluation is inseparable from the context in which data are generated.

Inspired by Evaluatology, we reflect on the current data-driven paradigm in AI. Regardless of whether observational or experimental, all data are inherently generated under specific conditions. However, prevailing AI training methods largely ignore these generative conditions and focus exclusively on the data themselves. Such a deficiency leads to uneven and difficult-to-evaluate data quality, constrains interpretability and the capacity for causal discovery, and renders models fragile in the face of novel scenarios.

Our research intuition is that explicitly incorporating both data and their generative conditions into the training process can substantially enhance the effectiveness and transparency of AI. Even under limited data availability, leveraging the interplay between data and conditions allows the discovery of deeper causal structures, enabling models to capture the invariant informational essence beneath data diversity. By grounding learning in condition-aware causal relationships, we move toward more robust, interpretable, and genuinely intelligent systems.

2.2. Motivation: The limitations of existing AI paradigm

The modern trajectory of data-driven artificial intelligence has been shaped by the belief that more data yields better models. This principle underlies the development of large language models (LLMs), whose performance scales predictably with training data volume, model size, and compute budget [26]. However, this scaling paradigm is increasingly constrained by a looming data bottleneck. As high-quality human-authored data becomes saturated and expensive to curate, synthetic data generation has emerged as a promising alternative.

Despite its scalability, synthetic data introduces a new layer of complexity [27–29]. Crucially, the quality of synthetic data is fundamentally limited by the generative models that produce it, which are often black-box architectures with little transparency or interpretability. This lack of visibility makes it difficult to trace the root causes of errors or biases in downstream models back to specific properties of the synthetic data. When performance deteriorates, it remains unclear whether the issue lies in data coverage, semantic consistency, or deeper representational flaws.

In practice, current synthetic data suffers from several well-documented issues: (1) generative models may fail to match the statistical distribution of real data, introducing biases that impair generalization. (2) synthetic samples often contain logical contradictions or distorted features that are difficult to detect but can corrupt pretraining (see Fig. 1). Low diversity and mode collapse: generators tend to produce samples with limited variation, leading to models that overfit narrow modes and underperform on real-world variability.

To improve the quality, reliability, and usefulness of synthetic data, it is imperative to enhance the interpretability and evaluation of generative models. Without understanding what a generator has learned — and what it systematically omits — scaling synthetic corpora becomes a blind process, susceptible to spurious correlations and misalignment.

These observations motivate a shift toward an evaluatology-driven AI paradigm, in which systematic attribution and interpretability are not afterthoughts but central components of the AI development cycle. By incorporating formal principles of evaluatology into the design, analysis, and deployment of generative models, we can better align synthetic data generation with downstream objectives, ensure quality control, and build AI systems that are not only larger, but measurably better.

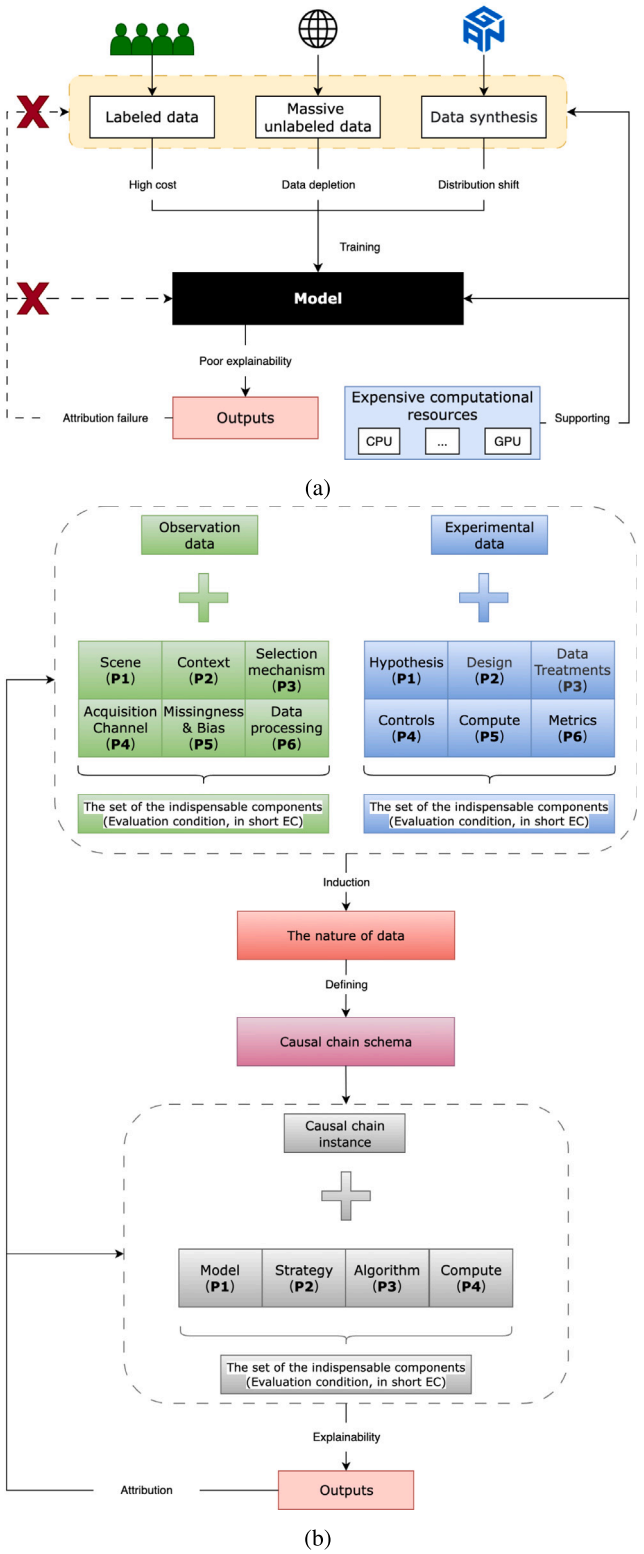


Fig. 1. Comparison of AI paradigms. (a) Data-driven: models are trained on massive, unstructured data and self-discover patterns, offering limited causal attribution and interpretability. (b) Evaluatology-driven: data are generated under explicit evaluation conditions within Self-contained Evaluation Systems and abstracted into causal-chain training instances, yielding interpretable, traceable, and attributable outputs.

3. The new AI paradigm based on evaluatology

3.1. Overview

Our methodology is grounded in *Evaluatology*, which reconceptualizes evaluation as the process of inferring the influence of an evaluated object on affected factors and attributing the observed outcomes to specific ones. The central methodological unit is the *Self-contained Evaluation System (SES)*, which anchors evaluation to its essential generative context.

Definition 3.1 (*Self-contained Evaluation System*). A Self-contained Evaluation System is defined as

$$SES = (E, C),$$

where E denotes the *evaluated object* and C denotes the *evaluation conditions*. The evaluation conditions C are composed of a minimal set of indispensable affected factors, which determine how variations in conditions propagate to outcomes.

Within this formalism, both data and models are situated in explicitly defined SESs. Data instances arise as functions of E and C, while induction over multiple condition configurations reveals their invariant informational essence. These essences can be abstracted into causal-chain schemas and instantiated into causal-chain instances, which serve as interpretable and reusable training data. Models themselves can also be represented as SESs, ensuring that their outputs are attributable to well-defined evaluation conditions. For example, in the task of video keyframe selection, event segmentation is one indispensable factor in C, and can be instantiated through visual-based shot boundary detection, audio-based segmentation via automatic speech recognition (ASR) pauses or speaker changes, or multimodal fusion of visual and auditory cues. The segmentation strategy determines how the video is partitioned and directly affects the structure of the output, illustrating how variations in condition configurations propagate to outcomes. Models themselves can also be represented as SESs, ensuring that their outputs are attributable to well-defined evaluation conditions. While defining an SES requires some domain-specific effort, it enforces strict evaluation conditions that enhance interpretability and reduce data dependency and long-term cost.

3.2. Self-contained evaluation systems for observational data

Observational data emerge from natural processes but are nonetheless shaped by indispensable generative factors as shown in Fig. 1(b). We formalize their essential affected factor set as

$$\mathcal{O} = \{S, T, M, A, B, P\},$$

where S is the *Scene*, T the *Context*, M the *Selection Mechanism*, A the *Acquisition Channel*, B the *Missingness & Bias*, and P the *Data Processing*.

Given an evaluated object E, an observational data instance d^{obs} is generated as

$$d^{obs} = f_{obs}(E, \mathcal{O}),$$

where f_{obs} denotes the observational generative mapping. By considering multiple admissible configurations $\mathcal{O}_i \in \Omega_{obs}$, we obtain the invariant informational essence of observational data:

$$\phi^{obs} = \text{Induction}\big(\{f_{obs}(E, \mathcal{O}_i) \mid \mathcal{O}_i \in \Omega_{obs}\}\big).$$

3.3. Self-contained evaluation systems for experimental data

Experimental data are deliberately generated under interventions and controls, guided by explicit causal inquiry. Their essential affected factor set is formalized as

$$\mathcal{X} = \{H, D, K, Q, R\},$$

where H is the *Hypothesis*, D the *Design*, K the *Controls*, Q the *Compute*, and R the *Metrics*.

For an evaluated object E, an experimental data instance d^{exp} is generated as

$$d^{exp} = f_{exp}(E, \mathcal{X}),$$

with f_{exp} representing the experimental generative mapping. The invariant informational essence of experimental data is induced as

$$\phi^{exp} = \text{Induction}\big(\{f_{exp}(E, \mathcal{X}_j) \mid \mathcal{X}_j \in \Omega_{exp}\}\big).$$

3.4. Induction and causal chain construction

Both ϕ^{obs} and ϕ^{exp} serve as abstract representations of the invariant essence of data. These are formalized as *causal-chain schemas* that describe how changes in evaluation conditions propagate to outcomes. When instantiated under concrete configurations of $\mathcal{O}$ or $\mathcal{X}$, these schemas yield *causal-chain instances*, which serve as interpretable, reusable, and condition-aware training examples.

3.5. Self-contained evaluation systems for models

Models themselves operate within SESs. The essential affected factor set for models can be denoted as

$$\mathcal{M} = \{T, A, Q, I\},$$

where T denotes the *Strategy*, A the *Algorithm*, Q the *Compute*, and I the *Implementation*.

The output of a model y can then be expressed as

$$y = f_{model}(E, \mathcal{M}),$$

with f_{model} representing the mapping from the evaluated object E and model-level evaluation conditions $\mathcal{M}$ to outputs. Since y is generated under explicit evaluation conditions, it is inherently explainable and attributable to specific elements of $\mathcal{M}$.

3.6. Case study: applying self-contained evaluation systems to video retrieval

Our experiments show that a user's query of vector database is aligned with only about **1.95%** of the keyframes in long-form videos. Motivated by an evaluation-based perspective, we introduce a *essential factor set* to defining the SES for keyframe selection — *event segmentation*, *textual signals*, *temporal context*, *scene/quality constraints*, *selection mechanism*, and *de-redundancy*. This essential factor set, grounded in a self-contained evaluation system, not only enables fine-grained attribution of retrieval performance to individual design factors, but also exemplifies the evaluatology-driven methodology underlying automatic database design.

In the Self-contained Evaluation System (SES), the **evaluation object** E is the target keyframe. The **evaluation conditions** C refer to a essential set of indispensable affected factors (see Definition 3.1):

1. Segment each video into *clips* using shot boundaries together with automatic speech recognition (ASR) pauses and speaker changes;
2. Derive a short *title/summary* per clip from subtitles/ASR as textual features;
3. Rank candidate frames by *textual match* (best matching 25 or embedding similarity) and by *temporal proximity* to query-relevant timestamps (closer is better);
4. Apply *scene/quality filters* (avoid blur and heavy motion; prefer stable frames shortly after shot transitions);
5. Within each clip, keep *1–3* frames and *merge near-duplicates* (retain a single frame for adjacent time spans);
6. Enforce a *keep-rate* of $\kappa \approx 2\%$–3% via a gating threshold, relaxing to $\sim 5\%$ when textual evidence is sparse.

By restricting both training and inference to keyframes identified by the SES, the model concentrates on *causally relevant semantics*, achieving retrieval quality comparable to full-frame pipelines while *substantially reducing* training data and inference cost.

3.7. Summary

In summary, our methodology situates both data and models within explicit Self-contained Evaluation Systems, each defined by an evaluated object E and its indispensable evaluation conditions C. By formalizing observational and experimental data generation as functions of (E, C), and by extending the same logic to models, we establish a unified framework in which invariant informational essences can be induced, abstracted into causal-chain schemas, and instantiated into training data. This paradigm transforms evaluation into an active methodological principle that spans the entire AI lifecycle, enabling interpretability, traceability, and epistemic grounding in artificial intelligence. Notably, this paradigm provides a theoretical foundation for automatic database design, by enabling the isolation of external confounding factors and attributing performance outcomes directly to the database design itself.

4. Conclusion

This work advances an evaluatology-based paradigm that spans data generation, training, and assessment. Its central construct — the Self-contained Evaluation System — couples an evaluated object with evaluation conditions constituted by a essential set of indispensable affected factors, thereby fixing the context required for coherent causal attribution. Within SESs, data are generated under explicit conditions, distilled into invariant informational structures, abstracted as causal-chain schemas, and instantiated as training examples; consequently, evaluation becomes a generative principle rather than a post-hoc procedure. The paradigm yields condition-aware learning with interpretable, traceable outputs and reduces dependence on massive, unstructured corpora. Taken together, these elements provide a scalable, unifying foundation for transparent, robust, and epistemically grounded AI, and furnish a precise basis on which future theory, benchmarks, and systems can be systematically developed.

CRediT authorship contribution statement

Guoxin Kang: Writing – original draft. **Wanling Gao:** Writing – review & editing. **Jianfeng Zhan:** Writing – review & editing, Conceptualization.

Funding

This paper is supported by the Strategic Research Special Funding of the Bureau of Development and Planning, Chinese Academy of Sciences (GHJ-ZLZX-2024-34).

Declaration of competing interest

The authors declare no competing interests.

References

[1] Y. Bengio, A. Courville, P. Vincent, Representation learning: A review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1798–1828.

[2] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, F.A. Wichmann, Shortcut learning in deep neural networks, Nat. Mach. Intell. 2 (11) (2020) 665–673.

[3] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D.d.L. Casas, L.A. Hendricks, J. Welbl, A. Clark, et al., Training compute-optimal large language models, 2022, arXiv preprint arXiv:2203.15556.

[4] P. Villalobos, A. Ho, J. Sevilla, T. Besiroglu, L. Heim, M. Hobbhahn, Will we run out of data? Limits of LLM scaling based on human-generated data, 2024, Epoch AI Blog, https://epoch.ai/blog/will-we-run-out-of-data-limits-of-llm-scaling-based-on-human-generated-data. (Accessed 28 August 2025).

[5] I. Shumailov, Z. Shumaylov, Y. Zhao, Y. Gal, N. Papernot, R. Anderson, The curse of recursion: Training on generated data makes models forget, 2023, arXiv preprint arXiv:2305.17493.

[6] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artificial Intelligence Res. 16 (2002) 321–357.

[7] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, P. Abbeel, Domain randomization for transferring deep neural networks from simulation to the real world, in: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, IEEE, 2017, pp. 23–30.

[8] L. Xu, M. Skoularidou, A. Cuesta-Infante, K. Veeramachaneni, Modeling tabular data using conditional gan, Adv. Neural Inf. Process. Syst. 32 (2019).

[9] J. Jordon, J. Yoon, M. Van Der Schaar, PATE-GAN: Generating synthetic data with differential privacy guarantees, in: International Conference on Learning Representations, 2018.

[10] S.R. Richter, V. Vineet, S. Roth, V. Koltun, Playing for data: Ground truth from computer games, in: European Conference on Computer Vision, Springer, 2016, pp. 102–118.

[11] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, et al., A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, ACM Trans. Inf. Syst. 43 (2) (2025) 1–55.

[12] J. Zhan, L. Wang, W. Gao, H. Li, C. Wang, Y. Huang, Y. Li, Z. Yang, G. Kang, C. Luo, et al., Evaluatology: The science and engineering of evaluation, BenchCouncil Trans. Benchmarks, Stand. Eval. 4 (1) (2024) 100162.

[13] J. Zhan, Fundamental concepts and methodologies in evaluatology, BenchCouncil Trans. Benchmarks, Stand. Eval. 4 (3) (2024) 100188.

[14] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst. 25 (2012).

[15] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, P. Kuksa, Natural language processing (almost) from scratch, J. Mach. Learn. Res. 12 (7) (2011).

[16] D. Rolnick, A. Veit, S. Belongie, N. Shavit, Deep learning is robust to massive label noise, 2017, arXiv preprint arXiv:1705.10694.

[17] S. Lloyd, Least squares quantization in PCM, IEEE Trans. Inform. Theory 28 (2) (1982) 129–137.

[18] I. Jolliffe, Principal component analysis, in: International Encyclopedia of Statistical Science, Springer, 2011, pp. 1094–1096.

[19] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 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), 2019, pp. 4171–4186.

[20] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: International Conference on Machine Learning, PmLR, 2020, pp. 1597–1607.

[21] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick, Masked autoencoders are scalable vision learners, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16000–16009.

[22] R.S. Sutton, A.G. Barto, et al., Reinforcement Learning: An Introduction, vol. 1, no. 1, MIT press Cambridge, 1998.

[23] Q. Yu, Z. Zhang, R. Zhu, Y. Yuan, X. Zuo, Y. Yue, W. Dai, T. Fan, G. Liu, L. Liu, et al., Dapo: An open-source llm reinforcement learning system at scale, 2025, arXiv preprint arXiv:2503.14476.

[24] C. Sun, S. Huang, D. Pompili, Llm-based multi-agent reinforcement learning: Current and future directions, 2024, arXiv preprint arXiv:2405.11106.

[25] C. Wang, L. Wang, W. Gao, Y. Yang, Y. Zhou, J. Zhan, Achieving consistent and comparable CPU evaluation outcomes, 2024, arXiv preprint arXiv:2411.08494.

[26] J. Kaplan, S. McCandlish, T. Henighan, T.B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, D. Amodei, Scaling laws for neural language models, 2020, arXiv preprint arXiv:2001.08361.

[27] I. Shumailov, Z. Shumaylov, Y. Zhao, N. Papernot, R. Anderson, Y. Gal, AI models collapse when trained on recursively generated data, Nature 631 (8022) (2024) 755–759.

[28] A. Mumuni, F. Mumuni, N.K. Gerrar, A survey of synthetic data augmentation methods in computer vision, 2024, arXiv preprint arXiv:2403.10075.

[29] A. Bauer, S. Trapp, M. Stenger, R. Leppich, S. Kounev, M. Leznik, K. Chard, I. Foster, Comprehensive exploration of synthetic data generation: A survey, 2024, arXiv preprint arXiv:2401.02524.