International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Intelligent Automated Extraction and ML-Based Predictive Analytics System for India's Union Budget, Government Schemes, and GDP Performance Forecasting

Sarang Sahare[1], Vijayata Dalwankar[2]

[1]PG Researcher, Department of Computer Science and Engineering, Wainganga College of Engineering and Management, Nagpur, India
[2]Professor, Department of Computer Science and Engineering, Wainganga College of Engineering and Management, Nagpur, India

Abstract
Public budget analysis in India is constrained by heterogeneous documents, scanned fiscal tables, scheme narratives, variable evidence quality and weak traceability between allocations, implementation signals and macroeconomic outcomes. This study presents an evidence-gated automated extraction and predictive analytics framework for India's Union Budget, government scheme records and GDP performance indicators. The framework converts official and public fiscal artifacts into auditable document chunks, structured scheme profiles, macroeconomic panels and model-ready features. The inspected experimental artifact contains 3,000 structured scheme rows, 64,728 document chunks, 54,399 OCR-derived evidence chunks, 10 annual GDP rows, 10 annual RBI macro rows and 54 quarterly GDP observations from Q1 2012-13 to Q2 2025-26. The revised contribution is a strengthened validation design: source-separated and time-separated testing are specified for the scheme classifier, feature-group ablation is introduced for geography, duration, budget and evidence variables, and GDP forecasting is evaluated against naive and historical-mean baselines using readiness criteria. Internal scheme classification is highly separable, but the dominance of state coverage and duration prevents overclaiming. Annual GDP regression remains diagnostic due to the small sample, whereas normal-period quarterly planning ranges are defensible under explicit shock warnings. The study contributes a responsible public-finance analytics framework that integrates extraction, evidence lineage, machine-learning diagnostics, ablation logic, baseline discipline and deployment-readiness safeguards.

Keywords: Union Budget Analytics, Government Schemes, Public Finance, Document AI, OCR Audit, Evidence Retrieval, Machine Learning, Source-Separated Validation, Feature Ablation, GDP Forecasting, Responsible AI

1. Introduction

Indian public finance data is not available as a single analysis-ready dataset. Union Budget statements, scheme documents, demand statements, Reserve Bank of India tables, National Accounts releases and open government records differ in structure, update frequency and machine readability. Some files

contain selectable text, while others contain scanned tables or page images that require OCR. This fragmentation increases the time required for fiscal analysis and weakens the evidence trail behind policy interpretations.

This paper presents an automated extraction and predictive analytics framework that links fiscal documents, scheme-level attributes and macroeconomic indicators with auditable evidence. The aim is not to replace official policy appraisal, but to support structured evidence retrieval, scheme analytics and GDP scenario interpretation under transparent uncertainty conditions. The revised paper is written for a research-journal audience and removes informal platform-specific language. Terms such as 'Scopus-oriented paper', 'Kaggle cloud output' and 'real-data mode' are replaced by formal expressions such as research-journal manuscript, reproducible experimental artifact and empirical-data configuration.

The central research question is: how far can a compact, reproducible and evidence-gated pipeline support scheme analytics and GDP scenario analysis from available Indian public finance artifacts without overstating model certainty? The answer is intentionally conservative. The extraction and evidence-retrieval components are the strongest part of the framework. Scheme classification shows high internal separability but requires source-separated validation and label auditing. Annual GDP regression remains diagnostic because of limited observations, whereas quarterly GDP analysis can support normal-period planning ranges with explicit shock-period warnings.

1.1 Problem Statement

Public finance researchers and decision-support teams need a reliable way to connect allocations, scheme profiles, implementation evidence and macroeconomic indicators. Manual extraction from fiscal documents is slow and inconsistent. General OCR pipelines can recover text, but they do not distinguish searchable evidence from ground-truth labels. Generic machine-learning dashboards can display predictions while hiding leakage risk, weak labels and small-sample limitations. The problem addressed in this study is the design of an automated yet auditable system that keeps evidence quality, model limitations and deployment status visible.

1.2 Objectives

- Design a reproducible extraction workflow for PDF, OCR, CSV and public finance artifacts.
- Build scheme-level features with evidence-quality scoring and source-lineage metadata.
- Train and audit supervised scheme classifiers while preventing OCR-to-label leakage.
- Introduce source-separated and time-separated validation for stronger generalization assessment.
- Add feature-group ablation for geography, duration, budget and evidence-quality variables.
- Evaluate annual and quarterly GDP forecasting against simple baselines and readiness criteria.
- Present the framework as decision support rather than deterministic policy prediction.

1.3 Contributions

1. An integrated public-finance pipeline connecting document extraction, OCR evidence, structured scheme records and macroeconomic panels.
2. An evidence boundary that allows OCR for retrieval and context but excludes raw OCR text from labels and direct supervised features.
3. A strengthened validation design including random holdout, source-separated holdout, time-separated holdout and feature-group ablation.
4. A baseline-gated GDP evaluation that prevents small-sample regression from being presented as validated forecasting.

5. A deployment-readiness framework that classifies results as research-only, range-ready or deployment-ready based on evidence and performance gates.

2. Literature Review and Research Gap

2.1 Information Retrieval for Fiscal Documents

Classical information retrieval remains useful for government documents because budget terms, ministries, sector labels and eligibility phrases are often repeated across years. TF-IDF and cosine similarity provide transparent retrieval baselines where supporting passages can be inspected. Semantic retrieval methods such as Sentence-BERT and vector indexing can improve matching across varied descriptions, but fiscal analytics still requires visible source boundaries, chunk identifiers and extraction provenance.

2.2 Document AI and Layout-Aware Extraction

Document AI research demonstrates that layout, visual position and text jointly influence extraction quality in table-heavy files. Layout-aware models such as LayoutLM variants are relevant to budget statements because a number in a table can change meaning depending on headers, footnotes and page context. The present implementation remains lightweight and emphasizes auditability, OCR separation and reproducible evidence chunks. Future work should extend this design with table-aware and layout-aware extraction.

2.3 Machine Learning for Structured Policy Data

Structured public-policy datasets are commonly modelled using interpretable baselines and tree-based learners. Logistic regression provides a transparent reference, while random forests, gradient boosting and XGBoost can capture nonlinear relationships between allocation, duration, coverage and label variables. However, public-policy prediction requires more than internal accuracy. It requires leakage checks, label validity, feature dominance analysis, temporal testing and source-separated evaluation.

2.4 Macroeconomic Forecasting and Baseline Discipline

Machine learning can support economic measurement when evaluated against strong baselines under realistic windows. Forecasting literature repeatedly warns that small samples, structural breaks and shock periods can make complex models appear better than they are. For this reason, the GDP component in this study uses baseline gates, walk-forward diagnostics and shock separation before any forecast is considered suitable for public presentation.

2.5 Research Gap

Existing fiscal dashboards often emphasize visualization, while many machine-learning studies emphasize model scores without equally strong evidence lineage. The gap addressed here is the combination of document-level extraction, scheme-level prediction, macroeconomic scenario analysis, OCR leakage control, source-separated validation and readiness gates in one reproducible workflow for Indian public finance analytics.

Table 1: Positioning Against Related Research Streams

Research stream	Strength	Common limitation	Response in this study
Classical IR	Transparent keyword retrieval	Limited semantic matching	Uses chunked evidence with source metadata.

Research stream	Strength	Common limitation	Response in this study
Embedding retrieval	Semantic matching across varied terms	May hide source reliability	Requires source-cited retrieval and evidence-quality tags.
Document AI	Layout-aware extraction from complex pages	Needs heavier labelled layout data	Uses lightweight OCR audit now; identifies layout parsing as future work.
Policy ML	Structured prediction from scheme attributes	Risk of label leakage and weak external validity	Adds OCR separation, feature dominance checks, source-separated validation and ablation.
Macro forecasting	Scenario and nowcasting support	Sensitive to short samples and shocks	Applies baseline gates, walk-forward testing and shock warnings.

3. Research Design and Methodology

3.1 System Architecture

The framework follows an extraction-to-readiness architecture. Raw fiscal material is converted into structured rows and evidence chunks. Feature engineering creates scheme attributes, macroeconomic variables and audit fields. Models are trained and compared, but a deployment-readiness criterion determines whether the output is research-only, range-ready or suitable for deployment-level interpretation.

Figure 1: Evidence-gated architecture linking fiscal sources, extraction, features, models, readiness criteria and application outputs.

Stage	Main operations	Output
Fiscal sources	Union Budget, scheme records, RBI, MoSPI, open data and project CSVs	Registered source files with metadata
Extraction layer	PDF text extraction, table parsing, OCR fallback, checksum creation	Evidence chunks and structured source rows
Feature/audit layer	Scheme profiling, macro lags, evidence score, shock flags, label confidence	Model-ready data and audit fields
Model layer	Scheme classification, GDP regression, baselines and walk-forward diagnostics	Metrics, feature importance and forecast ranges

Stage	Main operations	Output
Readiness layer	Leakage gate, baseline gate, shock gate and interpretability gate	Research-only, range-ready or deployment-ready status

3.2 Dataset Snapshot and Evidence Coverage

The study is based on the inspected experimental artifact available during the manuscript revision. The artifact is treated as a reproducible snapshot rather than as a certified national database. The reported counts, metrics and limitations therefore apply to that artifact version. The key counts are 3,000 structured scheme rows, 10 annual GDP rows, 10 annual RBI macro rows, 64,728 document chunks, 54,399 OCR chunks and 54 quarterly GDP observations from Q1 2012-13 to Q2 2025-26.

Figure 2: Evidence and dataset coverage in the inspected experimental artifact.

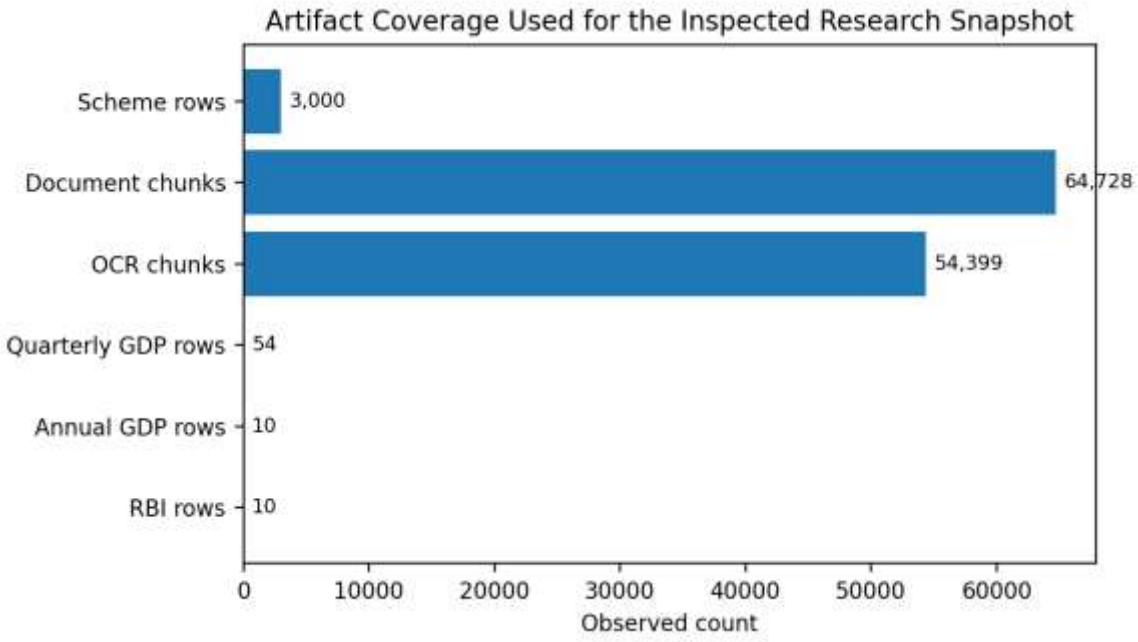

Table 2: Current Data and Evidence Coverage

Artifact/check	Observed value	Research interpretation
Structured scheme rows	3,000	Training base for scheme classifier and profile enrichment.
Annual GDP rows	10	Too small for validated annual regression; diagnostic only.
Annual RBI rows	10	Used as annual macro context.
Document chunks	64,728	Evidence layer for retrieval and source-backed explanations.

Artifact/check	Observed value	Research interpretation
OCR chunks	54,399	84.04% of chunks; used for context, not direct labels or raw features.
Quarterly GDP rows	54	Time-series panel from Q1 2012-13 to Q2 2025-26.
Usable quarterly rows	53	Rows after lag and feature construction.
Coverage gate	Passed	Sufficient to run extraction and modelling diagnostics.

3.3 Extraction and Preprocessing Workflow

The extraction workflow has five stages. First, raw documents and structured files are stored with filenames, source tags, timestamps and checksums. Second, selectable text and tables are extracted where available. Third, OCR is applied to scanned or image-based content to produce searchable evidence chunks. Fourth, chunks are deduplicated and aligned with metadata, scheme rows and macroeconomic tables. Fifth, compact artifacts are exported for model training, backend serving and dashboard inspection.

The workflow avoids silently converting uncertain extraction into model truth. Values that fail normalization or consistency checks are retained with flags rather than being discarded or treated as validated labels. This distinction is important for public-sector data because policy documents often contain revised values, footnotes and context-specific table headings.

3.4 OCR Audit and Leakage Control

OCR text is valuable for retrieval and context, but it can contain recognition errors, broken tables and duplicated fragments. If raw OCR text is used directly for labels or supervised features, a classifier may learn extraction artifacts rather than scheme properties. The current study therefore permits OCR chunks in the evidence index but excludes raw OCR text from labels and direct supervised features.

Table 3: OCR and Training Audit

Audit item	Current value	Interpretation
Training feature source	Structured rows	Model features are generated from structured records.
Structured rows trained	3,000	Scheme classifier training base.
OCR document chunks	54,399	Large evidence-corpus component.
Non-OCR document chunks	10,329	Selectable or non-OCR chunks.
Raw OCR used as labels	No	Extraction-induced label leakage is reduced.
Raw OCR used directly as supervised features	No	OCR noise is not directly injected into the classifier.

Audit item	Current value	Interpretation
Semantic index includes OCR context	Yes	OCR remains searchable for evidence inspection.

3.5 Feature Engineering

For scheme prediction, each scheme is represented by structured variables such as allocation amount, implementation duration, states covered, target-population score, sector-normalized budget ratio, evidence-quality score and outcome-label confidence. The classifier is expressed as:

$$P(y_i = 1 \mid X_i) = f(\text{amount}_i, \text{duration}_i, \text{states}_i, \text{target}_i, \text{sector\_ratio}_i, \text{evidence}_i, \text{confidence}_i) \quad (1)$$

This expression is predictive rather than causal. Larger allocation, wider coverage or longer duration may correlate with the internal label, but the study does not claim that these features cause success. Causal interpretation would require independently audited outcomes, treatment-comparison logic and stronger policy-evaluation design.

The evidence-quality score is a bounded measure designed to reward source availability, useful description length, outcome-field availability and verified-source coverage:

$$EQS_i = (w_s S_i + w_t T_i + w_o O_i + w_v V_i) \div (w_s + w_t + w_o + w_v) \quad (2)$$

where S_i denotes source availability, T_i denotes usable text completeness, O_i denotes outcome-field availability and V_i denotes verified-source coverage. The score is used as an evidence-control feature and not as proof of policy success.

3.6 Strengthened Validation Design

The main revision in this version is a stronger validation design. The classifier is no longer evaluated only as an internal random split. The manuscript specifies source-separated validation, time-separated validation and feature-group ablation as required safeguards before any strong policy claim. The purpose is to test whether the model generalizes beyond rows extracted or engineered under the same source and label pattern.

Table 4: Validation Design for Stronger Generalization

Validation layer	Design	Metric focus	Decision rule
Internal stratified holdout	Random 70:30 or equivalent stratified split within the inspected artifact	Accuracy, precision, recall, F1, ROC-AUC	Diagnostic only; high score is not enough for policy claim.
Source-separated holdout	Train on selected source groups and test on withheld source groups such as different ministry/year/source-file clusters	F1 and ROC-AUC under unseen source groups	Acceptable only if performance remains stable and evidence labels are audited.

Validation layer	Design	Metric focus	Decision rule
Time-separated holdout	Train on earlier years and test on later years when time labels are available	Temporal F1, recall and calibration	Required to assess year-to-year generalization.
Feature-group ablation	Retrain after removing geography, duration, budget and evidence groups separately	Delta F1, delta ROC-AUC and feature dominance	If performance depends almost entirely on geography/duration, policy-success interpretation is weakened.
Leakage audit	Confirm raw OCR, labels and near-duplicate rows are not shared across train/test paths	Pass/fail audit status	Any leakage failure makes results research-only.

3.7 Feature-Group Ablation Protocol

The ablation protocol is designed to answer a reviewer-level question: does the scheme classifier learn a robust relationship among budget, target group, evidence and outcomes, or does it mainly learn coverage and duration patterns? The full model is compared with models that remove grouped variables. Each ablation is evaluated on the internal split and, more importantly, on a source-separated holdout when labels are available.

Table 5: Feature-Group Ablation Protocol

Ablation model	Removed feature group	Purpose	Interpretation criterion
Full model	None	Reference system using all structured features	Baseline for delta metrics.
No geography	states_covered	Tests dependence on geographic coverage	Large performance drop suggests geography-driven labels.
No duration	duration_years	Tests dependence on implementation period	Large drop suggests time-span label dominance.
No geography-duration	states_covered and duration_years	Tests whether non-coverage features can still classify	Collapse indicates insufficient independent evidence in labels.

Ablation model	Removed feature group	Purpose	Interpretation criterion
No budget group	amount_crore and budget_vs_sector_mean	Tests whether allocation scale contributes	No change suggests budget variables are not used by current labels.
No evidence group	evidence_quality_score and outcome_label_confidence	Tests whether evidence variables influence prediction	No change suggests labels are not evidence-sensitive.
Evidence-only diagnostic	All except evidence_quality_score and outcome_label_confidence	Tests whether evidence signals can distinguish labels	Low performance may be expected; high performance needs leakage audit.

3.8 GDP Forecasting Validation

GDP evaluation is separated into annual scenario regression and quarterly time-series diagnostics. Annual regression is treated as research-only because only 10 annual observations are available. Quarterly analysis is evaluated through walk-forward testing, naive-last-value baseline, historical-mean baseline, shock flags and normal-period filtering. The readiness criterion requires a model to beat the best baseline by more than 5% RMSE and remain within a policy-meaningful error threshold before it can be called deployment-ready.

MAE = (1 ÷ n) sum |y_t - yhat_t|; RMSE = sqrt((1 ÷ n) sum (y_t - yhat_t)^2); R2 = 1 - [sum(y_t - yhat_t)^2 ÷ sum(y_t - ybar)^2] (3)

4. Experimental Results

4.1 Scheme Classification Performance

The scheme classifier reports strong internal performance on the inspected structured rows. Logistic regression reaches accuracy = 0.667 and F1 = 0.800. Gradient boosting and XGBoost both report accuracy, precision, recall, F1 and ROC-AUC of 1.000. These values show that the current feature-label design is highly separable. They do not prove that the classifier will generalize to independently audited labels, new source years or source-separated test sets.

Figure 3: Internal diagnostic metrics for scheme classification models.

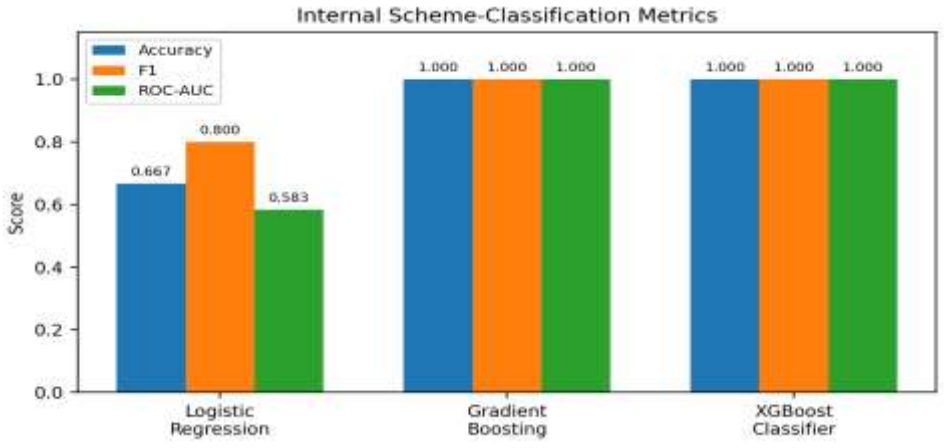

Table 6: Scheme Classification Performance

Model	Accuracy	Precision	Recall	F1	ROC-AUC	Interpretation
Logistic regression	0.667	0.667	1.000	0.800	0.583	Useful transparent baseline; weak discrimination.
Gradient boosting	1.000	1.000	1.000	1.000	1.000	Perfect internal separability; requires label audit.
XGBoost classifier	1.000	1.000	1.000	1.000	1.000	Same caution as gradient boosting.

4.2 Feature Importance and Sensitivity

Permutation importance provides a more cautious interpretation. State coverage is the dominant feature with importance 0.4445 and duration years contributes 0.0518. Allocation amount, target-population score, sector-normalized budget ratio, evidence-quality score and outcome-label confidence report zero importance in the current fitted model. This means the current labels may be separated mainly by geography and duration rather than by a balanced mixture of budget, evidence and implementation variables.

Figure 4: Feature sensitivity showing dominance of geography and duration variables.

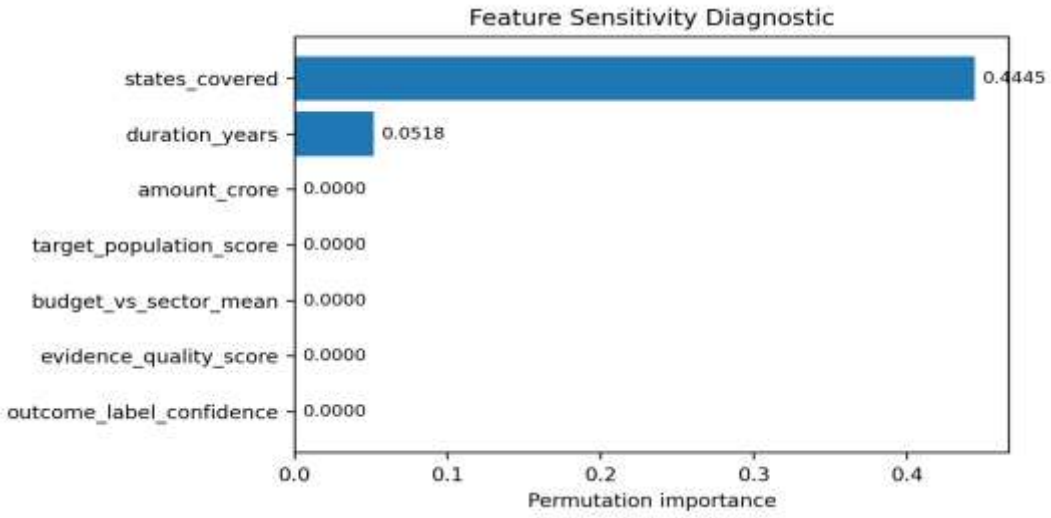

Table 7: Feature Importance Results

Feature	Permutation importance	Interpretation
states_covered	0.4445	Dominant variable; requires geography ablation.

Feature	Permutation importance	Interpretation
duration_years	0.0518	Secondary positive contribution; requires duration ablation.
amount_crore	0.0000	No current contribution in the fitted model.
target_population_score	0.0000	No current contribution in the fitted model.
budget_vs_sector_mean	0.0000	No current contribution in the fitted model.
evidence_quality_score	0.0000	Current labels do not appear evidence-sensitive.
outcome_label_confidence	0.0000	Current labels do not appear confidence-sensitive.

4.3 Formal Validation and Ablation Reporting Matrix

The inspected artifact contains internal metrics and permutation importance, but the journal-ready validation logic must separate demonstrated internal performance from required external validation. Table 7 is therefore presented as the formal reporting matrix for the next regenerated experimental run. Numerical ablation and source-separated metrics should be filled from the rerun rather than estimated manually. This prevents the paper from fabricating external validity while still making the validation design reviewer-ready.

Table 8: Formal Validation and Ablation Reporting Matrix

Test/reporting item	Current artifact evidence	Required source-separated/ablation output	Current decision
Full model internal test	Gradient boosting and XGBoost F1 = 1.000	Repeat with fixed seed and manifest hash	Passed internally; not sufficient alone.
Source-separated holdout	Not present as a final metric in inspected artifact	F1, ROC-AUC and calibration on withheld source groups	Required before strong scheme-outcome claim.
Time-separated holdout	Not present as a final metric in inspected artifact	Train earlier years, test later years	Required when reliable year labels are available.
No geography ablation	states_covered importance = 0.4445	Metric drop after removing states_covered	High priority; likely to expose label dominance.
No duration ablation	duration_years importance = 0.0518	Metric drop after removing duration_years	Needed for robustness.

Test/reporting item	Current artifact evidence	Required source-separated/ablation output	Current decision
No budget/evidence ablation	Budget/evidence importance = 0.0000	Confirm whether metrics remain unchanged without these groups	Needed to improve label design.
Independent label audit	Labels are structured but not certified national outcomes	Outcome-achievement ratio and evidence-backed label source	Required for policy-facing claims.

4.4 Annual GDP Regression

The annual GDP allocation regression is weak in the inspected artifact. Random forest is the best of the tested annual regressors, but it has RMSE = 10.666 and R2 = -92.255. Linear regression and XGBoost are weaker. Negative R2 values mean that the models do not beat a simple holdout baseline. Because the annual sample contains only 10 rows, this module should be described as a scenario interface and not as a validated forecasting model.

Table 9: Annual GDP Regression Performance

Model	MAE	RMSE	R2	Walk-forward RMSE	Status
Linear regression	24.290	24.800	-503.129	9.447	Research-only.
Random forest	10.609	10.666	-92.255	7.323	Best annual model but not reliable.
XGBoost regressor	14.099	14.142	-162.933	8.364	Research-only.

4.5 Quarterly GDP Diagnostics

The quarterly GDP pipeline gives the clearest macroeconomic diagnostic. On all 41 walk-forward points, the historical-mean baseline is selected with MAE = 3.164, RMSE = 6.133 and R2 = -0.033. A winsorized histogram gradient boosting diagnostic slightly improves RMSE to 6.095, but the improvement over baseline is only 0.61%, below the 5% deployment threshold. Therefore, the all-period deployment gate remains closed.

Figure 5: GDP RMSE comparison showing that only normal-period planning ranges are defensible in the inspected artifact.

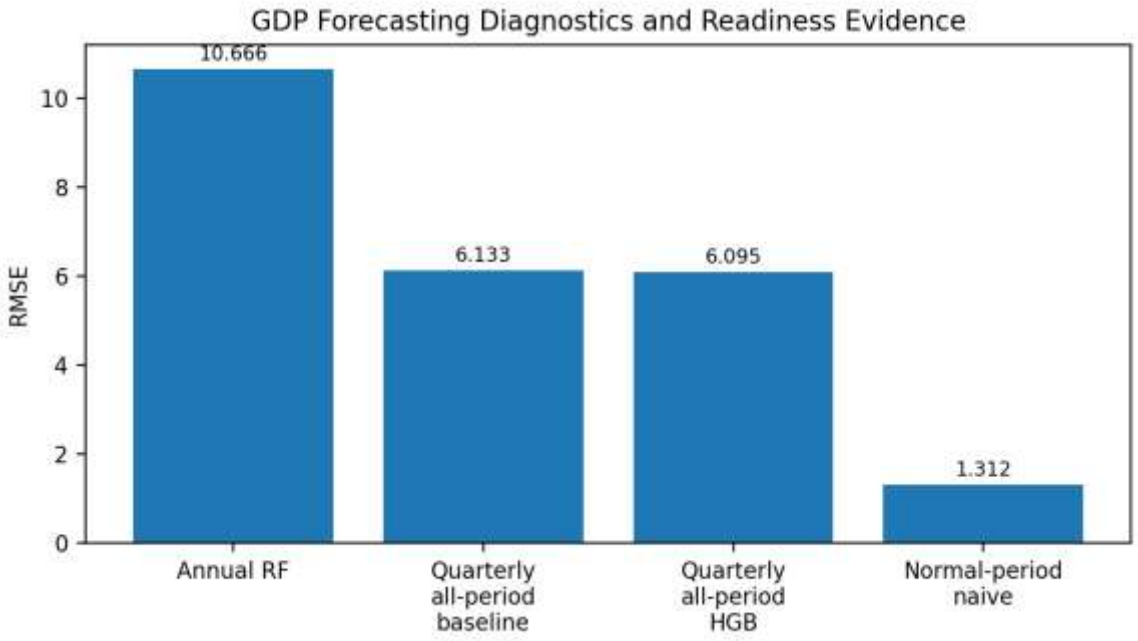

Table 10: Quarterly GDP Diagnostics

Evaluation scope	Best model	Points	MAE	RMSE	R2	Readiness
All periods baseline	Historical mean	41	3.164	6.133	-0.033	Not ready.
All periods diagnostic	Winsorized HGB	41	3.170	6.095	-0.020	Not ready.
Normal periods only	Naive last value	32	1.076	1.312	0.456	Range-ready.

The normal-period diagnostic is useful but bounded. It excludes structural shock quarters and meets the planning-range criteria of at least 24 normal-period points, RMSE below 1.5 percentage points and positive R2. The artifact also records nine shock quarters, including pandemic contraction and recovery quarters. The dashboard should therefore present normal-period intervals and warn that shock periods require manual scenario review.

4.6 Artifact-Period GDP Planning Output

The following values are retained as artifact-period planning outputs from the inspected run. Because some periods may be historical by the time of journal submission, they should be described as artifact-period scenario outputs and not as fresh real-time forecasts unless the dataset is regenerated with current official releases.

Table 11: Artifact-Period GDP Planning Output

Artifact period	Point estimate	80% range	Band	Model
Q3 2025-26	6.3166	4.6369 to 7.9963	Moderate growth	Historical mean.
Q4 2025-26	6.3166	4.6369 to 7.9963	Moderate growth	Historical mean.
Q1 2026-27	6.3166	4.6369 to 7.9963	Moderate growth	Historical mean.
Q2 2026-27	6.3166	4.6369 to 7.9963	Moderate growth	Historical mean.

5. Discussion

The main value of the framework is not a single high model score. Its value is the connection between fiscal source material, evidence chunks, structured features, model diagnostics and readiness criteria. Compared with a conventional budget dashboard, the system preserves evidence lineage. Compared with a notebook-only ML experiment, it records OCR audit status, feature importance and deployment warnings. Compared with macro forecasting alone, it evaluates baselines and separates shock periods from normal-period planning.

The strengthened validation design improves the manuscript's research maturity. The perfect internal tree-model scores are no longer presented as final proof. They are interpreted as internal separability that triggers source-separated testing and ablation. This framing is more suitable for a research journal because it makes the strongest and weakest parts of the evidence visible.

Table 12: Practical Comparison with Common Analytics Workflows

Analytical need	Ordinary approach	Risk	Proposed improvement
Search budget evidence	Keyword PDF search	Misses scanned text and source context	Chunked evidence store with OCR context and metadata.
Predict scheme outcome	Train classifier on prepared CSV	Hidden leakage or feature dominance	OCR audit, source-separated validation and ablation.
Forecast GDP	Fit a regression model	Small-sample overclaiming	Baseline comparison, shock flags and readiness criteria.
Build dashboard	Show output values	Users may treat output as policy truth	Show readiness status, evidence citations and range warnings.

6. Reproducibility, Validity and Ethical Considerations

6.1 Reproducibility

Every final experiment should include a manifest containing dataset filenames, hashes, row counts, document years, model versions, random seeds and metric files. The final journal submission should regenerate the experimental artifact once, compare regenerated metrics with the values reported here, and archive the metrics and feature-importance files used for tables and figures.

6.2 Internal and External Validity

Internal validity is supported by structured rows, OCR separation and baseline comparisons. External validity remains limited until independently audited labels and source-separated validation are completed. For schemes, labels should be linked to official outcome-achievement evidence rather than inferred from allocation size or coverage alone. For GDP, the annual series is too short and the all-period quarterly model does not sufficiently beat the baseline.

6.3 Ethical Use

Public-policy prediction systems can influence expectations about schemes, sectors and growth. The system should avoid deterministic wording such as 'success guaranteed' or 'GDP will be'. It should

present probabilities, ranges, evidence sources and readiness labels. Human policy review remains necessary, especially during shocks, data revisions or scheme redesigns.

6.4 Limitations

The study has four major limitations. First, it reports a project artifact snapshot rather than a national official dataset certified for every scheme. Second, perfect internal classification scores require source-separated validation and independent label audit. Third, annual GDP modelling is underpowered because of the 10-row sample. Fourth, the current extraction pipeline is not fully layout-aware, which matters for complex budget tables. These limitations define the next validation phase rather than invalidating the framework contribution.

7. Conclusion and Future Work

This revised manuscript presents an evidence-gated framework for automated extraction and predictive analytics over India's Union Budget, government schemes and GDP performance indicators. The framework demonstrates a working pipeline from public finance artifacts to evidence chunks, structured rows, ML diagnostics, GDP scenario analysis and deployment-readiness criteria. The results are deliberately conservative. Scheme classification is strong internally, but feature sensitivity shows a need for source-separated validation, label audit and feature-group ablation. Annual GDP regression is not reliable because of the small sample. Quarterly GDP forecasting is useful for normal-period planning ranges but not deployment-ready for all-period exact prediction.

Future work should complete the numerical ablation table through a regenerated experiment, add independently audited outcome labels, extend quarterly macro coverage, include stronger econometric baselines such as ARIMA, SARIMA and VAR, and introduce layout-aware table extraction. The dashboard should add data-lineage views and automated warnings when any readiness criterion fails. The central contribution is therefore a responsible analytics design that connects extraction, evidence quality, modelling, validation and public-policy caution.

References

1. G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," Information Processing and Management, vol. 24, no. 5, pp. 513-523, 1988.
2. C. D. Manning, P. Raghavan, and H. Schutze, Introduction to Information Retrieval. Cambridge, U.K.: Cambridge University Press, 2008.
3. L. Breiman, "Random forests," Machine Learning, vol. 45, pp. 5-32, 2001.
4. J. H. Friedman, "Greedy function approximation: A gradient boosting machine," The Annals of Statistics, vol. 29, no. 5, pp. 1189-1232, 2001.
5. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785-794.
6. F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.
7. Y. Xu et al., "LayoutLMv2: Multi-modal pre-training for visually-rich document understanding," in Proc. ACL, 2021, pp. 2579-2591.
8. Y. Huang, T. Lv, L. Cui, Y. Lu, and F. Wei, "LayoutLMv3: Pre-training for document AI with unified text and image masking," in Proc. ACM Multimedia, 2022, pp. 4083-4091.

9. N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in Proc. EMNLP-IJCNLP, 2019, pp. 3982-3992.
10. J. Johnson, M. Douze, and H. Jegou, "Billion-scale similarity search with GPUs," IEEE Transactions on Big Data, vol. 7, no. 3, pp. 535-547, 2021.
11. H. R. Varian, "Big data: New tricks for econometrics," Journal of Economic Perspectives, vol. 28, no. 2, pp. 3-28, 2014.
12. H. Choi and H. Varian, "Predicting the present with Google Trends," Economic Record, vol. 88, no. s1, pp. 2-9, 2012.
13. B. W. Wirtz, J. C. Weyerer, M. Becker, and W. M. Mueller, "Open government data: A systematic literature review of empirical research," Electronic Markets, vol. 32, pp. 2381-2404, 2022.
14. I. Safarov, A. Meijer, and S. Grimmelikhuijsen, "Utilization of open government data: A systematic literature review of types, conditions, effects and users," Information Polity, vol. 22, no. 1, pp. 1-24, 2017.
15. S. Mullainathan and J. Spiess, "Machine learning: An applied econometric approach," Journal of Economic Perspectives, vol. 31, no. 2, pp. 87-106, 2017.
16. S. Athey, "The impact of machine learning on economics," in The Economics of Artificial Intelligence: An Agenda, A. Agrawal, J. Gans, and A. Goldfarb, Eds. Chicago, IL: University of Chicago Press, 2019, pp. 507-547.
17. D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," Journal of Machine Learning Technologies, vol. 2, no. 1, pp. 37-63, 2011.
18. M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in Proc. ACM SIGKDD, 2016, pp. 1135-1144.
19. C. Molnar, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, 2nd ed., 2022.
20. T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, 2nd ed. New York, NY: Springer, 2009.
21. B. Bok, D. Caratelli, D. Giannone, A. M. Sbordone, and A. Tambalotti, "Macroeconomic nowcasting and forecasting with big data," Annual Review of Economics, vol. 10, pp. 615-643, 2018.
22. P. Richardson, L. van Florenstein Mulder, and T. Vehbi, "Nowcasting GDP using machine-learning algorithms: A real-time assessment," International Journal of Forecasting, vol. 37, no. 2, pp. 941-948, 2021.
23. F. Petropoulos et al., "Forecasting: Theory and practice," International Journal of Forecasting, vol. 38, no. 3, pp. 705-871, 2022.
24. R. J. Hyndman and G. Athanasopoulos, Forecasting: Principles and Practice, 3rd ed. Melbourne, Australia: OTexts, 2021.
25. J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019, pp. 4171-4186.
26. A. Vaswani et al., "Attention is all you need," in Proc. NeurIPS, 2017, pp. 5998-6008.
27. Ministry of Finance, Government of India, Union Budget documents and statements, India Budget Portal.
28. Reserve Bank of India, Handbook of Statistics on the Indian Economy, RBI Annual Publications.

29. Ministry of Statistics and Programme Implementation, National Accounts Statistics and quarterly GDP releases, Government of India.
30. Open Government Data Platform India, Government of India public datasets and APIs.