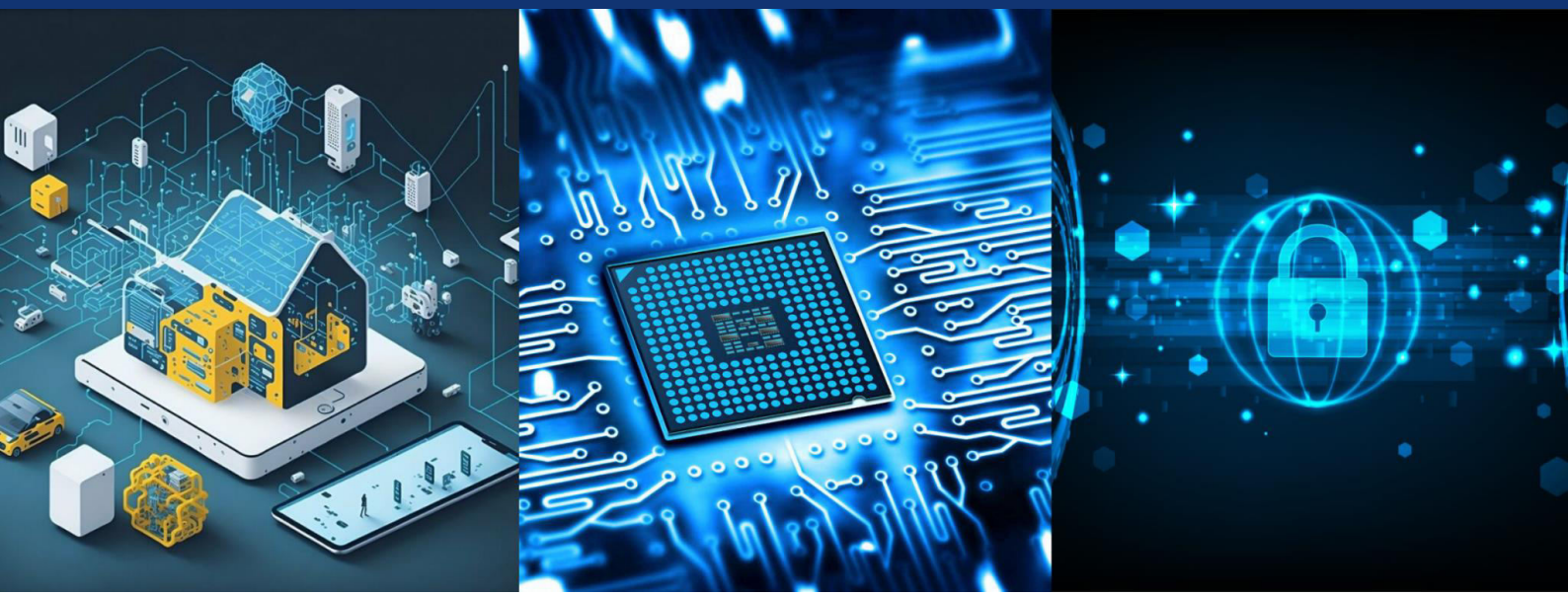
IJIRCCE

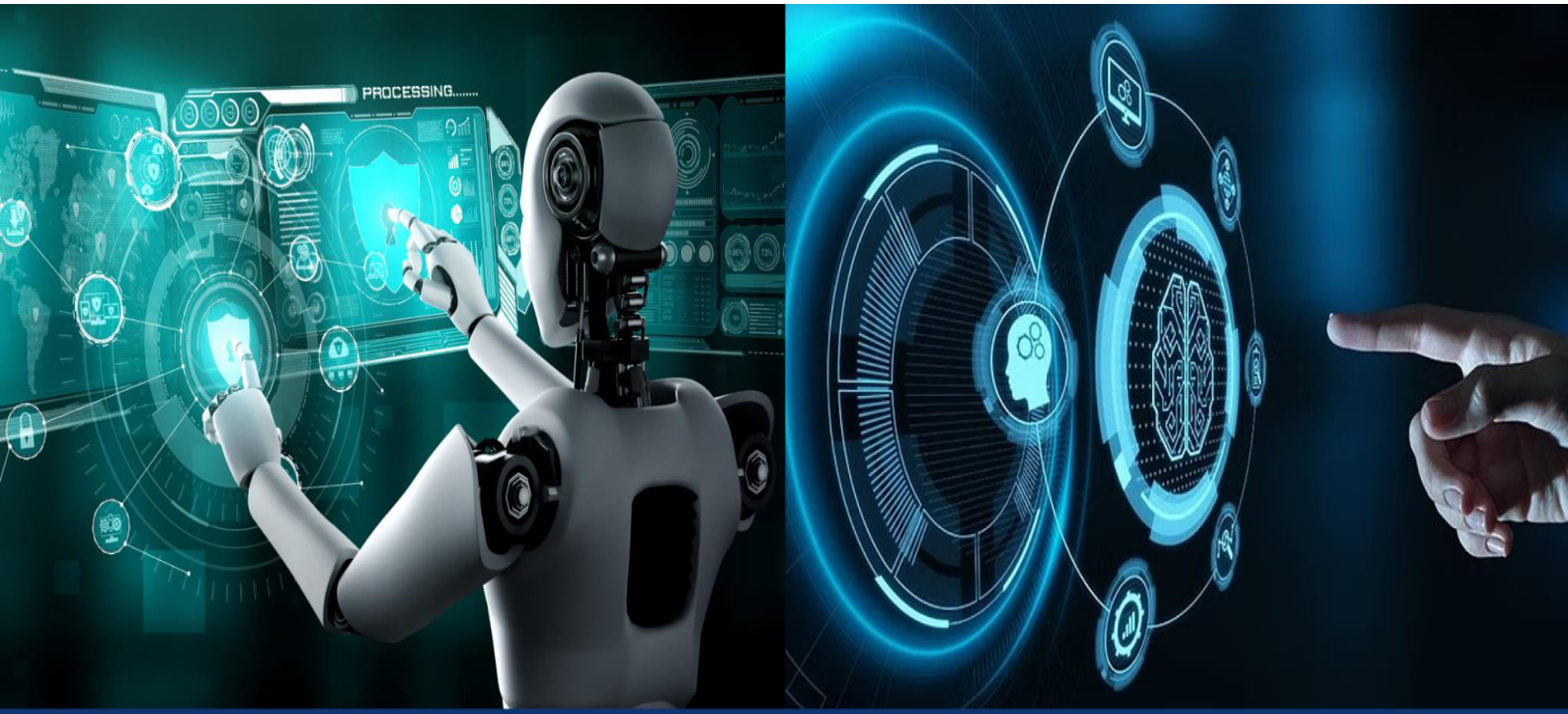
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 9.927

Volume 14, Issue 11, November 2026

© 2026 IJIRCCE | Volume 14, Issue 11, November 2026 DOI: 10.15680/IJIRCCE.2026.1411012

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

A Hybrid Ensemble Framework for Role-Sensitive Municipal Workforce-Efficiency Prediction: A Proxy-Dataset Benchmark and Validation Protocol for GHMC

C. Narahari Raghu Prasad[1*], Dr. Jaikumar M. Patil[2]

Research Scholar, Computer Science Engineering, Manasarovar Global University, Sehore (MP), India[1]

Professor, faculty of Engineering Department of Computer Science Engineering, Manasarovar Global University, Sehore (MP), India[2]

[*]Corresponding Author

Abstract: Municipal workforce analytics needs models that are accurate, resistant to leakage, sensitive to job role, interpretable, and usable under human supervision. This paper proposes a hybrid machine-learning framework for future validation at the Greater Hyderabad Municipal Corporation (GHMC) and tests its technical feasibility on two proxy HR datasets, neither of which contains GHMC records. The first dataset (Employee_AllInOne) has 3,000 records and 38 integrated variables; the second (HR_Analytics_India) has 5,000 records and 13 variables. Six individual classifiers and two ensembles were assessed with stratified cross-validation and an 80/20 holdout. On Employee_AllInOne, LightGBM had the best macro-F1 (0.5722) and MCC (0.5505); XGBoost and weighted soft voting had the highest accuracy (0.8417); regularised stacking had the best balanced accuracy (0.6236); and XGBoost had the best ROC-AUC (0.8991). On HR_Analytics_India, Extra Trees had the best macro-F1 (0.5096), accuracy (0.4970), and balanced accuracy (0.4972), while logistic regression had the best ROC-AUC (0.7165). Building an ensemble did not guarantee better results, and the preferred model depended on the dataset and the metric. Because the splits were random and stratified and no GHMC data were used, the results are a feasibility benchmark, not evidence of readiness for deployment. The GHMC contribution is a proposed validation protocol: role-normalised targets, employee-grouped or temporal validation, calibration, fairness auditing, explanation-stability analysis, privacy controls, and documented human review.

KEYWORDS: municipal workforce analytics; ensemble learning; people analytics; leakage control; model calibration; explainable AI; human-in-the-loop decision support

I. INTRODUCTION

GHMC is the local governing authority for the Greater Hyderabad area and is responsible for a broad set of civic services and administrative duties [1]. Urban bodies in India face growing demand, pressure on infrastructure, fiscal limits, and calls for accountability, so workforce performance cannot be judged by the number of tasks completed alone. It also depends on how well and how quickly work is done, how complaints are resolved, whether procedures and safety rules are followed, how staff work together, and how public money is used [2].

Public-sector workforce decisions have traditionally rested on establishment records, seniority, attendance, and supervisor evaluations. These sources say little about how employee characteristics, workload, training, field conditions, and service quality relate to one another. HR analytics addresses this by turning data from several systems into evidence for planning and development, but it pays off only when data quality, analytical skill, organisational process, and the ability to act on findings develop together [3], [4]. More data does not automatically improve decisions, and predictive systems bring concerns about data quality, fairness, interpretability, privacy, and the repetition of historical inequities [5], [6]. A municipal model should therefore explain its output, separate controllable factors from structural ones, and keep meaningful human oversight.

© 2026 IJIRCCE | Volume 14, Issue 11, November 2026 DOI: 10.15680/IJIRCCE.2026.1411012

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Efficiency also has to be measured against each department's operating conditions. Sanitation, engineering, public health, and revenue roles differ in output targets, workload, and seasonality, and a single raw productivity measure favours whoever has the most countable outputs. Machine learning has been applied to municipal worker productivity [7], employee performance scores [8], and attrition [9], [10], and tree-based and ensemble methods tend to work well on mixed numerical and categorical records. None of this work settles whether such models transfer to a municipal workforce.

This paper keeps two things apart. The executed part is a proxy-dataset feasibility benchmark: eight models on two public-style HR datasets, evaluated with random stratified splits. The proposed part is a GHMC validation protocol with five components: (1) a multidimensional, role-normalised Workforce Efficiency Score (WES); (2) diverse base learners combined through leakage-resistant stacking; (3) employee-grouped and temporal validation; (4) explainability, fairness, privacy, and audit controls; and (5) supportive interventions such as training, workload review, or process improvement, with no automated disciplinary action. The benchmark shows that the pipeline runs and how model choice varies. It does not show that GHMC employees' efficiency can be predicted, and the framework is not a universal scoring instrument.

II. BACKGROUND AND RELATED WORK

2.1 Municipal efficiency and public-sector people analytics

Municipal efficiency is the relationship among resources, working conditions, employee ability, and the quality of service reaching citizens. It should not reward speed at the expense of quality, safety, legal compliance, or citizen satisfaction, so it is measured along six dimensions: output, timeliness, quality (rework, errors, reopened complaints), reliability, compliance and safety, and citizen-service outcome. A capability-opportunity-motivation logic organises the drivers: employees need the skills for the task, the tools and conditions to carry it out, and clear goals, recognition, and fair process. Treating efficiency as a fixed personal trait is the main risk this logic guards against.

Public agencies pursue equity, legality, employee rights, and public trust, not profit. A statistically sound model can still fail if it invites gaming, ignores differences in assignment difficulty, or shifts attention from system constraints to individuals, which is why public personnel analytics has to be read in its institutional context [2]. Records are usually spread across establishment, payroll, attendance, training, work-order, grievance, and inspection systems with different identifiers, update cycles, and definitions (a "closed" grievance may be resolved, disposed of administratively, duplicated, or transferred). Procedural justice matters as well: staff accept analytical support more readily when the purpose is clear, measures are job-relevant, data can be corrected, and decisions can be explained and contested.

2.2 Machine learning for workforce outcomes

Workforce forecasting covers performance, productivity, absenteeism, attrition, and engagement. Statistical and machine-learning models have been fitted to municipal worker productivity [7], performance scores [8], and attrition, where feature selection and explanatory analysis proved useful [9], [10]. These outcomes are not interchangeable: attrition is an event, absenteeism is a pattern with several causes, and productivity is often reduced to a few outputs. Most studies also report accuracy or AUC after a random split. When one employee contributes several records, the same person can appear in both training and test data, and the model can learn identity-linked patterns instead of general ones. Target leakage, meaning a feature recorded after the outcome such as a final review or a later disciplinary action, is a second threat. The protocol therefore treats employee-grouped and temporal validation as a minimum requirement.
Actionability is a further concern. Age or background may raise statistical accuracy but gives no legitimate basis for managerial action, whereas training completion, role-skill alignment, case allocation, equipment downtime, vacancy pressure, and supervisor span can be acted on. The framework sorts features into four classes: actionable, contextual, audit-only, and prohibited.

2.3 Ensembles and hybrid models

No algorithm is best across workforce datasets. Linear models are transparent but miss nonlinear structure; trees capture interactions and mixed data types; bagging and boosting combine models in different ways [11], [12]. Random forests add feature importance and resistance to overfitting [13], gradient boosting corrects residual error sequentially [14], XGBoost adds regularisation and scalability [15], and LightGBM speeds up training on large tabular data [16]. Stacking trains a meta-learner on base-model predictions [17], and it helps only when the base learners make different errors. That argument motivates a hybrid design mixing an interpretable model with higher-capacity learners.

© 2026 IJIRCCE | Volume 14, Issue 11, November 2026 DOI: 10.15680/IJIRCCE.2026.1411012

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

2.4 Explainability, fairness, and accountability

Ensemble complexity calls for both global explanations (which features drive the model overall) and local ones (why this prediction). SHAP gives additive contributions for both purposes [18], and LIME approximates the model locally [19]. Explanations should also be judged on whether they match documented feature meanings, stay stable under resampling, and separate individual signals from workload and organisational constraints, and the choice of method depends on the user and the decision at stake [20]. Fairness is not achieved by dropping protected attributes: proxies, target bias, uneven measurement, and historical allocation patterns can all produce disparities [21]. Groups, job families, and departments should be compared on error rates and calibration, and small samples should be disclosed, not summarised with unreliable numbers. Model cards [22] and datasheets [23] document purpose, limits, and data provenance, and both feed internal audit [24].

Taken together, the literature supports five design choices: multidimensional, normalised efficiency; data covering capability, workload, support, and service outcome; diverse ensembles with explicit leakage control; calibrated probabilities with confidence intervals; and explanations paired with fairness, privacy, and human-review checks. Its gap is the lack of an integrated framework meeting all five for a large Indian municipal corporation. Existing work leans on private-sector or turnover data, uses generic performance labels, reports accuracy without calibration or external validity, and rarely says what an organisation should do with a prediction.

III. PROPOSED GHMC FRAMEWORK

Everything in Sections 3 and 4, including employee-grouped and temporal validation, is a proposed design for the future GHMC study; none of it was executed in the proxy-dataset benchmark reported in Section 5. Figure 1 shows the proposed system and its functional components. Governance spans all stages. Data flows left to right from source systems through target and feature construction, the hybrid ensemble, and evaluation and explanation to human review, and a monitoring layer feeds outcomes, appeals, and drift findings back into earlier stages. Protected attributes reach only the fairness audit.

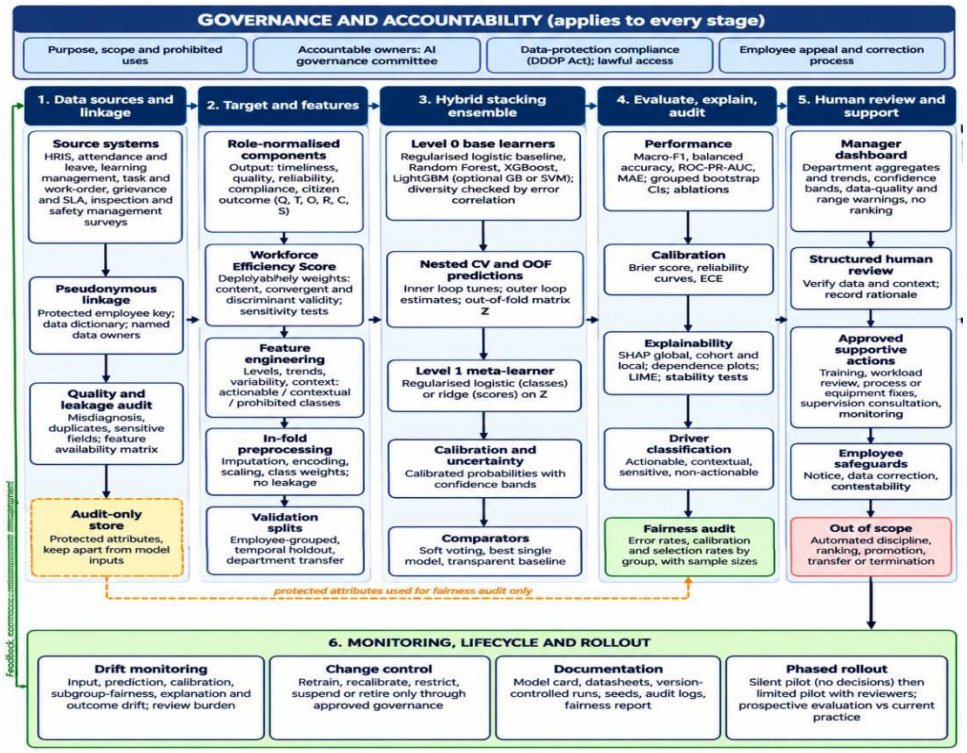


Figure.1. Proposed system architecture: governance, data and linkage, target and features, hybrid stacking ensemble, evaluation and explanation, human review, and monitoring.

© 2026 IJIRCCE | Volume 14, Issue 11, November 2026 DOI: 10.15680/IJIRCCE.2026.1411012

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

3.1 Constructs and hypotheses

Efficiency is treated as the outcome of capability (job knowledge, skill, experience, learning access), opportunity (resources, staffing, information, authority, equipment, coordination), motivation (goal clarity, recognition, fairness), and service context (demand, complexity, geography, seasonality, emergencies, dependence on other units). The model estimates the probability that an employee-period record falls in an efficiency class, or a continuous score. A prediction is anticipated performance, conditional on what was documented at prediction time, and it changes when the documented situation changes. Department and role information set expectations and never a lower standard for any group. Five directional hypotheses guide the GHMC study. H1: job-relevant training, skill alignment, and experience are positively associated with efficiency. H2: attendance regularity, timely completion, and less avoidable rework are positively associated. H3: excessive workload, unresolved constraints, and repeated resource shortages are negatively associated. H4: a diverse stacked ensemble generalises and calibrates better than any single base learner. H5: explanation analysis reveals organisational as well as individual drivers, supporting interventions beyond the employee. The baseline model tests average direction, trees and boosting can find thresholds and interactions, and H4 is tested against each component, a simple average, and a reduced ensemble.

3.2 Target construction: the Workforce Efficiency Score

Each role-normalised component (output O, timeliness T, quality Q, reliability R, compliance and safety C, citizen-service outcome S) is scaled to 0 to 100 for each job role and period and combined as

$$WES = w_O \cdot O + w_T \cdot T + w_Q \cdot Q + w_R \cdot R + w_C \cdot C + w_S \cdot S \qquad (1)$$

with non-negative weights summing to 1, set by analytic hierarchy process, Delphi consultation, or an approved performance framework, and tested by sensitivity analysis. Positive indicators are scaled by percentile as in Eq. (2), negative indicators such as rework are reversed, and extreme values are winsorised under a documented rule with originals kept for audit. Completed tasks are weighted by complexity band or expected effort.

$$O^* = 100 \times (O - P5_{role}) / (P95_{role} - P5_{role}), \text{ bounded to } [0, 100] \qquad (2)$$

A Delphi panel of HR staff, department managers, field and service representatives, quality experts, and data specialists agrees indicator definitions and weights first. Statistically derived weights (principal components or entropy) can be compared with expert weights but must not override service priorities. WES may be converted to three ordered classes with policy-approved or distribution-aware thresholds, and the continuous score is retained. Validation covers content validity, convergent validity against audited service quality, discriminant validity (WES should not merely repeat attendance or a supervisor rating), and temporal reliability. Experts review cases at both ends of the distribution against the underlying records. Target construction is finished before comparative modelling starts.

3.3 Data, features, and leakage control

Sources include HR information, attendance and leave, learning management, departmental task systems, grievance platforms, service-level monitoring, inspection and safety logs, and engagement surveys. Each has a named owner, lawful purpose, retention period, and quality assessment, and linkage runs through a protected pseudonymous key in line with data-protection law [25]. A data dictionary is fixed before modelling so that terms such as "completed task" are defined consistently. Table 1 summarises the variable groups.

© 2026 IJIRCCE | Volume 14, Issue 11, November 2026 DOI: 10.15680/IJIRCCE.2026.1411012

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Table.1. Proposed variable groups for the GHMC workforce-efficiency model

Variable group	Illustrative variables	Purpose and safeguards
Capability and experience	Job and role tenure, certified skills, training hours and completion	Job-relevant capability; unrelated personal qualifications excluded
Attendance and reliability	Attendance rate, approved leave, shift adherence, absence trend	Lawful leave never treated as misconduct
Task and service delivery	Assigned and completed tasks, on-time completion, backlog, grievance closure	Normalised by role, complexity, ward, season, resources
Quality and compliance	Rework, repeat complaints, audit observations, safety compliance	Validated records; individual versus team or system causes
Work environment	Team size, vacancy rate, supervisor span, equipment downtime, ward workload	Captures organisational constraints; prevents individual-only attribution
Engagement and support	Role clarity, supervisor support, tool adequacy, learning access	Voluntary, validated, aggregated survey measures

A feature enters the model only if it is relevant to the outcome, available before the prediction cut-off (timing), reliably recorded, and legitimate under its approved use. Direct identifiers, biometric templates, personal communications, unrelated medical records, and political or religious information are excluded, and audit attributes are stored separately. Features describe levels (current attendance, backlog, quality rate), trends against the employee's own baseline or the department, variability (instability in completion time), and context (vacancies, downtime, seasonal demand, dependencies). Rolling windows must end before the forecast window opens; a feature built from the forecast quarter's own outcome is leakage. A feature-availability matrix records the earliest legitimate timestamp for every variable. In the GHMC study, records will be grouped by employee, a temporal holdout will be kept for forecasting, and imputation, scaling, encoding, feature selection, and resampling will all be fitted inside training folds using scikit-learn pipelines [26]. Missing data are analysed by source, role, department, and period before imputation, and outliers are checked to see whether they are errors, emergency work, or genuine high performance. For imbalance, class weighting is the default; SMOTE [27] may be tested inside training folds only and must not be read as evidence about real employees [28]. Thresholds are set on validation data, from the cost of missed support needs, the cost of false alarms, and review capacity, and are frozen before the test set is touched.

3.4 Hybrid stacking architecture

Level-0 learners are a regularised linear or logistic model, random forest, XGBoost, and LightGBM, with gradient boosting or a support vector machine added only if it increases diversity. Diversity is judged by error correlation and disagreement, not by algorithm names; if two boosters make nearly identical errors, the more interpretable or better-calibrated one is kept. Each base learner is trained on K−1 folds and produces out-of-fold predictions for the held-out fold, giving a matrix Z used to train the meta-learner g:

$$z_{im} = f_m(-k(i))(x_i), \quad \hat{y}_i = g(z_{i1}, z_{i2}, \ldots, z_{iM}) \quad (3)$$

Here z holds class probabilities for classification or predicted scores for regression. The meta-learner is a regularised logistic model (classes) or ridge regression (scores), so its coefficients show each base model's weight, and it is never trained on in-sample predictions, which would inflate meta-level fit. Non-negative or simplex constraints can be tried for interpretability. Hyperparameters are tuned inside nested cross-validation (the inner loop chooses transformations and settings, the outer loop estimates generalisation) with tuning space, seeds, and software versions reported, and the test set is used once. Tuning prioritises macro-F1 and mean absolute error while tracking calibration and subgroup results. Ablations remove contextual, training, attendance, and support variables in turn, and remove calibration, imbalance handling, or the meta-learner. An unnormalised role target is kept only as a diagnostic baseline.

3.5 Evaluation, explainability, fairness, and oversight

Classification reports accuracy, balanced accuracy, macro precision and recall, macro-F1, weighted-F1, class-wise sensitivity and specificity, confusion matrices, ROC-AUC, and PR-AUC, since precision-recall curves are more informative when a class is rare [29]. Regression reports MAE, RMSE, R-squared, and error distributions by department. Calibration uses the Brier score, reliability plots, and expected calibration error. Confidence intervals come from grouped

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

bootstrap resampling, paired comparisons use the Wilcoxon signed-rank test or a bootstrap with an effect size, and multi-seed results check stability.

Explanations come at four levels: global importance, dependence plots, cohort summaries for similar roles, and local case displays showing the prediction, uncertainty, recent trend, main positive and negative contributors, and context warnings, each traceable to a source record. They are not causal findings, because a SHAP contribution describes how a feature moved the model relative to a reference distribution, not what would happen if the feature changed. Stability is checked by comparing feature rankings across bootstrap models or small perturbations, and instability triggers reviewer caution or model simplification. Fairness audits compare false-positive and false-negative rates, calibration, mean residuals, and selection rates across legally reviewed groups (such as gender, age band, disability status, employment group, job family, department, field versus office, and location), always with sample sizes and uncertainty. Protected attributes are used only in a separate audit dataset. A disparity leads to diagnosis (target definition, task complexity, missing context, data entry, real differences) and not to automatic metric equalisation, and no single fairness metric justifies deployment. Human review is structured and documented: reviewers see the prediction and explanation, confirm source data and context, choose from approved supportive actions, and record their reasoning. Employees are told when model output leads to an action and can correct data or add context, and appeals feed data-quality monitoring, never performance ratings.

IV. ALGORITHM AND IMPLEMENTATION ROADMAP

Algorithm 1 gives the full procedure proposed for the GHMC study; its employee-grouped and temporal validation steps have not yet been executed. It should run as a version-controlled pipeline that records the data snapshot, feature dictionary, target version, fold assignment, preprocessing objects, hyperparameters, metrics, explanations, and approval status. Test labels stay hidden from developers until preprocessing, feature selection, and model choices are frozen. Because employee records cannot be shared openly, reproducibility rests on transparent transformation logic, provenance, and auditable execution logs.

Implementation runs in six phases. Phase 1 fixes the decisions the model may inform, prohibited data uses, oversight roles, employee representation, legal review, and success measures. Phase 2 audits source systems, builds the data dictionary, tests linkage, designs the WES, and runs expert case review. Phase 3 builds the retrospective model with grouped and temporal validation. Phase 4 is a silent pilot in which predictions inform no decision and only expose data and workflow problems. Phase 5 is a limited pilot with authorised reviewers and a few approved interventions, measuring service outcomes, review burden, employee feedback, fairness, and unintended effects against current practice. Phase 6 decides whether to continue, revise, or retire the model. The pilot should cover one field service, one technical service, and one office or citizen-facing service, observed for 24 to 36 months where definitions are stable so that seasonal cycles are captured. The manager dashboard shows department-level aggregates and role trends first, not employee rankings, with confidence bands, out-of-range warnings, data-completeness flags, and a neutral, explained "support needed" label in place of unexplained red-amber-green codes.

Proxy-Dataset Benchmark

© 2026 IJIRCCE | Volume 14, Issue 11, November 2026 DOI: 10.15680/IJIRCCE.2026.1411012

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Algorithm.1. Hybrid ensemble workforce-efficiency procedure

1. Confirm approved purpose, scope, ownership, access rights, and prohibited uses for the authorised pseudonymised data.
2. Audit provenance, completeness, duplicates, target consistency, sensitive fields, and possible leakage.
3. Build comparable employee-period records and role-normalised O, T, Q, R, C, and S components; compute the continuous WES and ordered classes only if needed.
4. Split by employee group and time; keep an untouched temporal or departmental test set.
5. Build leakage-free preprocessing pipelines (missing values, encoding, scaling, feature filtering, imbalance handling).
6. Train the linear or logistic baseline, random forest, gradient boosting, XGBoost, and LightGBM under nested validation; generate out-of-fold outputs from the well-validated, diverse learners.
7. Train the regularised meta-learner on the out-of-fold outputs (Eq. 3).
8. Evaluate discrimination, error, calibration, stability, subgroup fairness, robustness, and operational usefulness.
9. Generate global, cohort, and local explanations; label each driver actionable, contextual, sensitive, or non-actionable.
10. Route eligible cases to human review; record the intervention, rationale, employee feedback, and outcome.
11. Monitor input and prediction drift, calibration, subgroup fairness, explanation stability, review burden, and intervention effect; retrain, recalibrate, restrict, suspend, or retire the model only through approved change control.
12. Return the model, model card, dataset documentation, fairness and audit report, review workflow, and dashboard.

4.1 Scope and setup

The benchmark evaluated six individual algorithms and two ensembles (weighted soft voting and regularised stacking) on two independent proxy datasets, reporting cross-validation summaries, holdout metrics, class-wise results, confusion matrices, and feature importance. It is preliminary for municipal use: validation used stratified random splits, not employee-grouped or temporal holdouts; repeated-fold intervals are missing for the ensembles; and there was no paired significance testing, calibration analysis, explanation-stability test, subgroup fairness audit, or GHMC validation. Table 2 gives the configuration.

Table.2. Dataset integration and experimental configuration

Dataset	Records	Input variables	Target	Classes	Test records	Best macro-F1 model
Employee_AllInOne	3,000	38 after merging	Performance score	4	600	LightGBM
HR_Analytics_India	5,000	13	rmance score in tertiles	3	1,000	Extra Trees

The first experiment used 3,000 integrated engagement and training records with 38 variables. Fourteen unusable, sensitive, or leakage-prone fields were excluded, leaving 23 raw predictors. The second used 5,000 India-based HR records with 13 variables; two unusable columns were removed and the continuous performance measure was converted

© 2026 IJIRCCE | Volume 14, Issue 11, November 2026 DOI: 10.15680/IJIRCCE.2026.1411012

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

to Low, Moderate, and High. The datasets differ in feature space, class structure, and target construction, so they are compared descriptively and not as a head-to-head benchmark.

4.2 Employee_AllInOne

Table.3. Performance on Employee_AllInOne (CV macro-F1 is three-fold mean ± SD; other columns are holdout)

Model	CV macro-F1	Accuracy	Bal. acc.	Macro-F1	MCC	ROC-AUC
LightGBM	0.4573 ± 0.0242	0.8333	0.5781	0.5722	0.5505	0.8947
Regularised Stacking	n/a	0.7150	0.6236	0.4961	0.4491	0.8975
Weighted Soft Voting	n/a	0.8417	0.4214	0.4491	0.4910	0.8696
XGBoost	0.3807 ± 0.0150	0.8417	0.3709	0.3848	0.4682	0.8991
CatBoost	0.3015 ± 0.0074	0.5133	0.3971	0.3078	0.1917	0.7113
Extra Trees	0.2847 ± 0.0101	0.5650	0.3187	0.2892	0.1234	0.6798
Random Forest	0.2672 ± 0.0110	0.6983	0.3017	0.2818	0.1639	0.7077
Logistic Regression	0.2695 ± 0.0026	0.5950	0.2520	0.2519	0.0129	0.6161

LightGBM had the strongest holdout macro-F1. Its accuracy, balanced accuracy, macro precision, macro recall, macro-F1, weighted-F1, MCC, and ROC-AUC were 0.8333, 0.5781, 0.5670, 0.5781, 0.5722, 0.8361, 0.5505, and 0.8947. Its cross-validated macro-F1 was 0.4573 ± 0.0242. Holdout performance above the CV mean suggests that the holdout split was easier than the average fold, another reason to repeat evaluation under grouped or temporal splits. XGBoost and weighted soft voting had a slightly higher accuracy (0.8417) but much lower macro-F1 (0.3848 and 0.4491), so accuracy alone favoured the majority class. Regularised stacking had the highest balanced accuracy (0.6236) and a ROC-AUC of 0.8975 but a macro-F1 of 0.4961, 0.0761 below LightGBM. XGBoost had the highest ROC-AUC (0.8991) and macro precision (0.6473) but a balanced accuracy of only 0.3709 and a macro-F1 of 0.3848. Ranking quality, class sensitivity, and default-threshold decisions therefore need separate evaluation.

4.3 HR_Analytics_India

Table.4. Performance on HR_Analytics_India (CV macro-F1 is mean ± SD; other columns are holdout)

Model	CV macro-F1	Accuracy	Bal. acc.	Macro-F1	MCC	ROC-AUC
Extra Trees	0.4801 ± 0.0044	0.4970	0.4972	0.5096	0.2482	0.7052
CatBoost	0.4712 ± 0.0043	0.4890	0.4893	0.4984	0.2355	0.7117
Weighted Soft Voting	n/a	0.4860	0.4864	0.4983	0.2311	0.7114
Regularised Stacking	n/a	0.4900	0.4906	0.4920	0.2348	0.7115
LightGBM	0.4709 ± 0.0204	0.4870	0.4879	0.4919	0.2309	0.6990
Logistic Regression	0.4792 ± 0.0131	0.4840	0.4847	0.4876	0.2258	0.7165
Random Forest	0.4834 ± 0.0103	0.4780	0.4784	0.4870	0.2177	0.7020
XGBoost	0.4771 ± 0.0019	0.4760	0.4764	0.4858	0.2154	0.7065

Extra Trees had the highest macro-F1 (0.5096), with accuracy 0.4970, balanced accuracy 0.4972, macro precision 0.5487, macro recall 0.4972, MCC 0.2482, and ROC-AUC 0.7052. Its cross-validated macro-F1 of 0.4801 ± 0.0044 is close to its holdout score, unlike on the first dataset. The other seven models scored between 0.4858 and 0.4984 on macro-F1, so their practical differences were small. Logistic regression had the highest ROC-AUC (0.7165), again showing that the preferred model depends on the objective.

© 2026 IJIRCCE | Volume 14, Issue 11, November 2026 DOI: 10.15680/IJIRCCE.2026.1411012

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

4.4 Cross-dataset and class-wise comparison

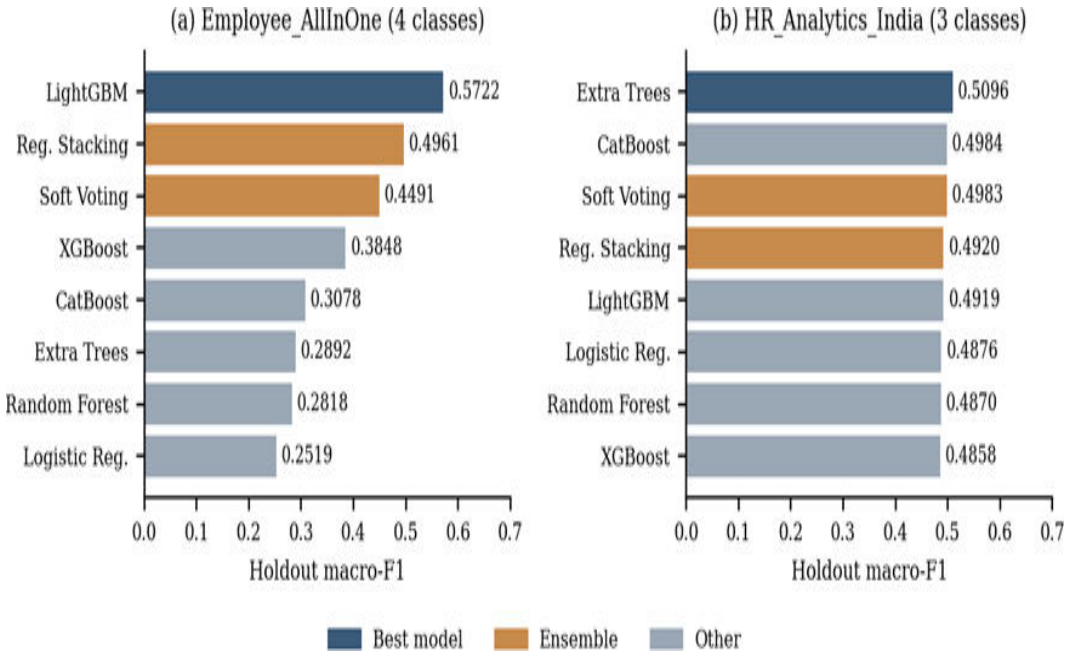


Figure.2. Holdout macro-F1 by model for (a) Employee_AllInOne and (b) R_Analytics_India.

Figure 2 shows that the best model depended on the dataset: LightGBM was clearly ahead on Employee_AllInOne and Extra Trees only marginally ahead on HR_Analytics_India. Ensembles did not guarantee higher macro-F1, which matters because stacking pays off only when base learners contribute complementary, well-calibrated errors. LightGBM also reached higher accuracy and ROC-AUC on Employee_AllInOne than Extra Trees did on HR_Analytics_India, but the datasets differ in class count, imbalance, features, and target construction, so this does not mean one is intrinsically more predictable.

Table.5. Class-wise results of the best model for each dataset

Dataset	Class	Precision	Recall	F1-score	Support
Employee_AllInOne	Exceeds	0.52	0.57	0.54	74
Employee_AllInOne	Fully Meets	0.94	0.92	0.93	472
Employee_AllInOne	Needs Improvement	0.39	0.40	0.39	35
Employee_AllInOne	PIP	0.42	0.42	0.42	19
HR_Analytics_India	High	0.82	0.52	0.64	327
HR_Analytics_India	Low	0.42	0.49	0.45	339
HR_Analytics_India	Moderate	0.41	0.49	0.44	334

LightGBM handled the dominant Fully Meets class well (precision 0.94, recall 0.92, F1 0.93) but was weaker on Exceeds (F1 0.54), Needs Improvement (0.39), and PIP (0.42), so high accuracy partly reflects the majority class and does not establish reliable detection of minority categories. On HR_Analytics_India the High class reached F1 0.64 while Low and Moderate reached 0.45 and 0.44; that similarity points to heavy overlap near the thresholds used to cut the continuous score into three classes. These results support class-specific reporting and argue against high-stakes use of the proxy models.

4.5 Discussion

The benchmark shows that heterogeneous HR records can be integrated, screened for sensitive or leakage-prone fields, turned into features, evaluated across bagging and boosting models, combined by voting or stacking, and assessed with

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

class-specific metrics. LightGBM led on one dataset and Extra Trees on the other, and neither ensemble beat the best single learner on macro-F1 on either dataset (stacking led only on balanced accuracy for Employee_AllInOne), which challenges the assumption that a more complex ensemble is better. Boosting and randomised tree methods look promising for heterogeneous workforce data, but the value of ensembling depends on the dataset. The evidence is enough for a pilot benchmark and not for claims of deployment readiness or of stacking's superiority. No employment decision should rest on these models. The next stage should replace random stratified splits with employee-grouped or temporal holdouts, repeat runs across seeds or folds, quantify uncertainty, audit target leakage, evaluate calibration, and test explanation stability, subgroup fairness, and the usefulness of support interventions. The protocol in Sections 3 and 4 builds these in from the start.

V. LIMITATIONS AND FUTURE RESEARCH

Five limitations bound the results. Neither proxy dataset contains GHMC records, so nothing here establishes performance in a municipal workforce. The two datasets use different targets and class structures, which limits direct comparison. Random stratified splitting does not test generalisation to unseen employees or future periods. Ensemble confidence intervals, paired significance tests, calibration curves, explanation-stability measures, and subgroup fairness audits are unavailable from the retained outputs. And the WES and the intervention workflow are proposals that still need GHMC-specific content validation, data-governance approval, and prospective evaluation.

Future work could use hierarchical models of workers nested in teams and wards, survival models for time to decline, and causal methods for intervention effects. Sequence models should be tried only if they beat simpler trend features and sample size and interpretability allow. Federated or other privacy-preserving learning could let municipalities collaborate without pooling raw employee data. Interviews and focus groups should run alongside the quantitative work to uncover unrecorded tasks, measurement burdens, and what staff consider efficient work, and comparisons across municipalities could test whether common feature families transfer while local targets stay local.

VI. CONCLUSION

This paper proposes a hybrid ensemble framework for municipal workforce-efficiency prediction and tests its technical feasibility on two proxy HR datasets. The results depend on the dataset and the metric: LightGBM had the best macro-F1 on Employee_AllInOne, Extra Trees on HR_Analytics_India, and neither ensemble beat the best individual learner on macro-F1. These findings support comparing diverse models but do not show that stacking is superior or that any model is ready for GHMC. The contribution for GHMC is the validation and governance protocol: role-normalised targets, leakage-resistant preprocessing, grouped or temporal testing, calibration, explainability, fairness auditing, privacy controls, human review, contestability, and monitoring. Workforce prediction here is decision support for finding training, workload, process, and resource constraints, not a basis for ranking or discipline. Only after GHMC-specific validation and a monitored silent pilot should the framework be considered for limited operational use, and its value will depend on reliable data, valid service standards, analytical skill, employee trust, and willingness to act on the organisational constraints the analysis reveals.

REFERENCES

1. Greater Hyderabad Municipal Corporation, "About GHMC," Official GHMC Portal, Hyderabad, India. Accessed: Jun. 20, 2026.
2. W. Cho, S. Choi, and H. Choi, "Human resources analytics for public personnel management: Concepts, cases, and caveats," Adm. Sci., vol. 13, no. 2, Art. no. 41, 2023, doi: 10.3390/admsci13020041.
3. J. H. Marler and J. W. Boudreau, "An evidence-based review of HR analytics," Int. J. Hum. Resour. Manag., vol. 28, no. 1, pp. 3–26, 2017, doi: 10.1080/09585192.2016.1244699.
4. D. Minbaeva, "Building credible human capital analytics for organizational competitive advantage," Hum. Resour. Manag., vol. 57, no. 3, pp. 701–713, 2018, doi: 10.1002/hrm.21848
5. A. Tursunbayeva, S. Di Lauro, and C. Pagliari, "People analytics: A scoping review of conceptual boundaries and value propositions," Int. J. Inf. Manag., vol. 43, pp. 224–247, 2018, doi: 10.1016/j.ijinfomgt.2018.08.002.
6. P. Tambe, P. Cappelli, and V. Yakubovich, "Artificial intelligence in human resources management: Challen ges and a path forward," Calif. Manag. Rev., vol. 61, no. 4, pp. 15–42, 2019, doi: 10.1177/0008125619867910.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

7. P. Bijalwan, A. Gupta, A. Mendiratta, A. Johri, and M. Asif, "Predicting the productivity of municipality workers: A comparison of six machine learning algorithms," Economies, vol. 12, no. 1, Art. no. 16, 2024, doi: 10.3390/economies12010016.
8. L. G. Tanasescu, A. Vines, A. R. Bologa, and O. Vîrgolici, "Data analytics for optimizing and predicting employee performance," Appl. Sci., vol. 14, no. 8, Art. no. 3254, 2024, doi: 10.3390/app14083254.
9. F. Fallucchi, M. Coladangelo, R. Giuliano, and E. W. De Luca, "Predicting employee attrition using machine learning techniques," Computers, vol. 9, no. 4, Art. no. 86, 2020, doi: 10.3390/computers9040086.
10. P. K. Jain, M. Jain, and R. Pamula, "Explaining and predicting employees' attrition: A machine learning approach," SN Appl. Sci., vol. 2, no. 4, Art. no. 757, 2020, doi: 10.1007/s42452-020-2519-4.
11. T. G. Dietterich, "Ensemble methods in machine learning," in Multiple Classifier Systems, Lecture Notes in Computer Science, vol. 1857. Berlin, Germany: Springer, 2000, pp. 1–15, doi: 10.1007/3-540-45014-9_1.
12. O. Sagi and L. Rokach, "Ensemble learning: A survey," WIREs Data Min. Knowl. Discov., vol. 8, no. 4, Art. no. e1249, 2018, doi: 10.1002/widm.1249.
13. L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
14. J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Ann. Statist., vol. 29, no. 5, pp. 1189–1232, 2001, doi: 10.1214/aos/1013203451.
15. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.
16. G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in Adv. Neural Inf. Process. Syst. 30, 2017, pp. 3146–3154.
17. D. H. Wolpert, "Stacked generalization," Neural Netw., vol. 5, no. 2, pp. 241–259, 1992, doi: 10.1016/S0893-6080(05)80023-1.
18. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Adv. Neural Inf. Process. Syst. 30, 2017, pp. 4765–4774.
19. M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.
20. A. Barredo Arrieta et al., "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," Inf. Fusion, vol. 58, pp. 82–115, 2020, doi: 10.1016/j.inffus.2019.12.012.
21. N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," ACM Comput. Surv., vol. 54, no. 6, Art. no. 115, 2021, doi: 10.1145/3457607.
22. M. Mitchell et al., "Model cards for model reporting," in Proc. Conf. Fairness, Accountability, Transparency, 2019, pp. 220–229, doi: 10.1145/3287560.3287596.
23. T. Gebru et al., "Datasheets for datasets," Commun. ACM, vol. 64, no. 12, pp. 86 –92, Dec. 2021, doi: 10.1145/3458723.
24. I. D. Raji et al., "Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing," in Proc. 2020 Conf. Fairness, Accountability, Transparency, 2020, pp. 33–44, doi: 10.1145/3351095.3372873.
25. Government of India, The Digital Personal Data Protection Act, 2023, Act No. 22 of 2023. New Delhi, India: Gazette of India, Aug. 11, 2023.
26. F. Pedregosa et al., "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011.
27. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.
28. H. He and E. A. Garcia, "Learning from imbalanced data," IEEE Trans. Knowl. Data Eng., vol. 21, no. 9, pp. 1263–1284, Sep. 2009, doi: 10.1109/TKDE.2008.239.
29. T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," PLOS ONE, vol. 10, no. 3, Art. no. e0118432, 2015, doi: 10.1371/journal.pone.0118432.



INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details