IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

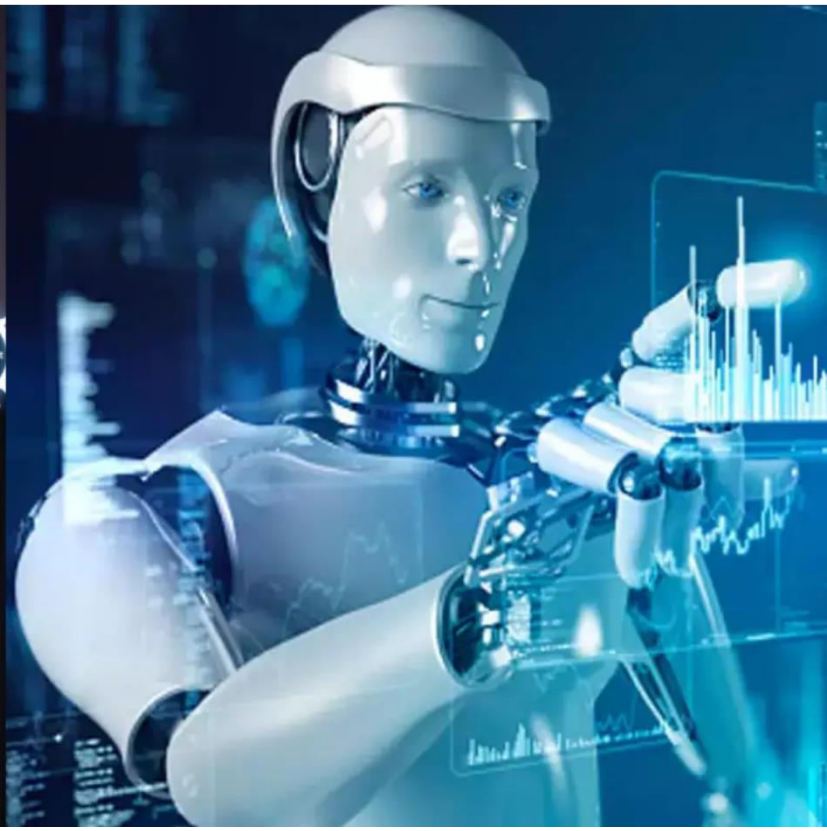
ISSN(O): 2319-8753
ISSN(P): 2347-6710



International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

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





Impact Factor: 9.935

Volume 15, Issue 11, November 2026

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 11, November 2026

|DOI: 10.15680/IJIRSET.2026.1511041|

Comparative Evaluation of Bias Detection and Mitigation Techniques in Classical Machine Learning and Large Language Models

Shilpashree V, Dr. Ramesh Sengodan

M. Tech Student, Department of Artificial Intelligence, Presidency University, Bengaluru, Karnataka, India

Professor, PSAIAC, Department of Artificial intelligence, Presidency University, Bengaluru, Karnataka, India

ABSTRACT: Algorithmic fairness assessment requires comparing predictive discrimination with group-conditioned outcome and error disparities across heterogeneous tasks. This research examines classical tabular classification on Adult Income, German Credit, and COMPAS, alongside transformer-based profession classification using Bias in Bios. Logistic Regression and XGBoost are evaluated with pre-processing, in-processing, and post-processing interventions including Reweighing, Disparate Impact Remover, Learning Fair Representations, Exponentiated Gradient Reduction, Reject Option Classification, Equalized Odds Postprocessing, and Calibrated Equalized Odds. The NLP comparison considers DistilBERT, GPT-2, BLOOM and BERT representation pathways paired with downstream classifiers. A fairness-audit application integrates FastAPI, LIME token attribution, CSV cohort analysis, and local language-model explanations. The recorded profession-audit comparison provides four signed Statistical Parity Difference values, with BERT–XGBoost under Exponentiated Gradient attaining the smallest absolute SPD among the reported configurations. Interpretation separates selection-rate parity from conditional error fairness, protected-group support, and predictive utility.

KEYWORDS: Algorithmic fairness, Statistical Parity Difference, Disparate Impact, Equal Opportunity Difference, Bias in Bios, Adult Income, COMPAS, XGBoost, LIME, mitigation.

I. INTRODUCTION

Predictive models increasingly influence decisions concerning income status, credit risk, recidivism, and occupational categorization. Training observations may encode historically unequal outcomes, demographic correlations, and proxy attributes. Consequently, high aggregate accuracy does not establish equitable model behavior across protected groups [1], [8]. A valid fairness audit must identify the favorable outcome, comparison groups, subgroup support, evaluation partition, and prediction threshold before interpreting disparity values.

This study covers four distinct benchmark tasks. Adult Income estimates a binary income threshold; German Credit represents credit-risk status; COMPAS relates to recidivism outcomes; and Bias in Bios assigns occupations from textual biographies [20]–[24]. A parity value calculated for one task cannot automatically be compared with a value computed under another target definition. The principal objective is thus to retain task-specific semantics while testing comparable mitigation stages.

Two modeling families are employed: classical Logistic Regression and XGBoost for structured tabular records, and transformer-derived text representations with downstream Logistic Regression or XGBoost for profession prediction. Fairness-aware interventions modify data weights, latent features, learning constraints, or post-hoc decisions. A FastAPI interface supports individual occupation prediction and batch fairness assessment without conflating local word attribution with population-level disparity.

II. RELATED WORK

Bellamy et al. established AIF360 as a common toolkit for fairness assessment and mitigation [8]. Kamiran and Calders investigated pre-processing to change discriminatory training distributions [12], whereas Zemel et al. studied fairness-aware latent representations [13]. Agarwal et al. formulated constrained classification as a sequence of cost-sensitive

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 11, November 2026

|DOI: 10.15680/IJIRSET.2026.1511041|

learning problems [9]. Hardt et al. developed equality-of-opportunity criteria based on error-rate behavior [10], and Pleiss et al. characterized tensions between calibration and fairness [11].

Research on text representations highlights the distinction between occupation prediction quality and encoded demographic associations. The Bias in Bios study examined representational gender bias in occupational classification [20]. BERT, DistilBERT, GPT-2, and BLOOM provide different tokenization and embedding mechanisms, meaning that the downstream classification and mitigation conditions should retain backbone-specific provenance [16]–[19]. LIME supplies local perturbation-based explanations but does not measure cohort fairness by itself [14].

Recent work also emphasizes instability caused by sampling, measurement assumptions, and imbalanced protected subgroups [1]–[7]. A fairness method may improve one parity criterion at the expense of classification performance or another fairness criterion. Model selection therefore requires simultaneous inspection of selection-rate and conditional-error measures, including the sign of the disparity, instead of treating a near-zero SPD as proof of universal fairness.

III. MATERIALS AND METHODS

A) Dataset and Target Construction

Adult Income, German Credit, and COMPAS are prepared through dataset-specific feature encoding, label orientation, protected-group mapping, and supervised partitioning. The tabular branch fits categorical encoders and imputers only on permitted training observations. Bias in Bios contains occupation-labeled biography text with gender-associated group identifiers. A multiclass profession audit must identify a selected positive profession or an explicit one-vs-rest evaluation; group fairness statistics are not defined by treating every occupation as the same favorable class.

B) Proposed Fairness-Aware Architecture

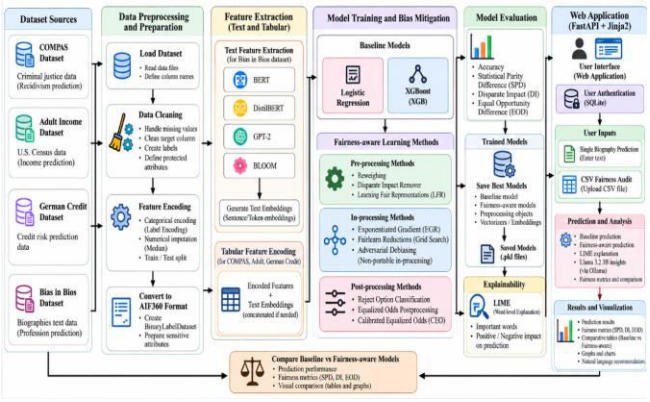


Fig.1. Data preparation, fairness mitigation, classification, and FastAPI audit architecture.

The architecture separates baseline model training from fairness intervention, evaluation, and web inference. The structured-data path uses Logistic Regression and XGBoost with Reweighing, Disparate Impact Remover, and Learning Fair Representations. Exponentiated Gradient Reduction constrains model fitting, while Reject Option Classification, Equalized Odds Postprocessing, and Calibrated Equalized Odds modify decisions after the base classifier. Results retain stage identity, sensitive-group encoding, and positive-outcome mapping.

C) Transformer-Based Profession Classifiers

The language-processing branch extracts embeddings using DistilBERT, GPT-2, BLOOM, or BERT and trains corresponding occupation predictors. Its stated configurations include DistilBERT–Logistic Regression, GPT-2–Logistic Regression, BLOOM–Logistic Regression, and BERT–XGBoost. The application emphasizes BERT pooled embeddings with XGBoost. Tokenizer versions, embedding dimensions, pooling rules, and label orders are fixed between training and inference. Gender-linked lexical proxies are inspected by LIME, while group differences are computed over labeled audit cohorts.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 11, November 2026

|DOI: 10.15680/IJIRSET.2026.1511041|

D) Mitigation and Controlled Evaluation

Reweighing assigns weights to observed protected-group and target combinations. Feature repair and representation learning modify the input distribution, whereas Exponentiated Gradient optimizes a fairness-constrained classifier mixture. Equalized Odds and Calibrated Equalized Odds use labeled score information to adjust a specified conditional-error objective; Reject Option Classification changes decisions near a confidence boundary. Each intervention requires a separately recorded calibration or training partition, and no mitigation tuning is performed on the final test observations.

E) Performance and Fairness Equations

For binary evaluation, let Y denote the observed favorable class, $\hat{Y}$ the predicted favorable class, and $A \in \{p,u\}$ the privileged/unprivileged group indicator. Accuracy measures aggregate correctness; SPD and DI compare positive decision rates, whereas EOD compares true-positive rates conditional on Y=1. Their signed or ratio interpretation depends on explicitly declared group ordering.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$SPD = \frac{N+(u)}{N(u)} - \frac{N+(p)}{N(p)} \quad (2)$$

$$DI = \frac{R(u)}{R(p)} \quad (3)$$

$$EOD = \frac{TP(u)}{TP(u)+FN(u)} - \frac{TP(p)}{TP(p)+FN(p)} \quad (4)$$

Let R(g) denote the favorable-prediction rate for protected group g. SPD measures R(u) minus R(p), DI is the ratio R(u)/R(p), and EOD measures the difference in true-positive rates. Group support and favorable-outcome definitions are required for each calculation. Sparse positive-label cells can make EOD undefined, and very small R(p) values can make DI unstable.

F) FastAPI Auditing and Explanations

The application authenticates users through SQLite-backed account workflows and provides two analysis paths. Single-biography inference generates an occupational label and word-level LIME explanation. Batch CSV analysis validates biography, profession, and gender fields before computing cohort-level parity and error statistics. A local Ollama Llama 3.2 3B service supplies explanatory wording, but deterministic model outputs and subgroup counts remain the authoritative basis of the assessment.

G) Dataset-Specific Class Semantics

Adult Income uses an income-threshold target whose positive state must be explicitly fixed before measuring favorable-decision parity. German Credit contains a credit-risk category, so the meaning of a favorable label depends on how the original class coding is converted. COMPAS instead relates to recidivism outcomes and error asymmetries in a justice-system context. These domains differ in legal and ethical interpretation despite sharing a binary classification interface. A common evaluator can compute rates, but it must not collapse these label definitions into a universal measure of substantive justice. The evaluation manifest therefore records the positive label, protected attribute, category ordering, and subgroup counts for each tabular benchmark.

Bias in Bios extends this framework to multiclass profession labels rather than the binary outcomes used by the structured benchmarks. A meaningful SPD for profession prediction requires either a declared one-vs-rest occupational class or a clearly specified aggregation rule. The audit also records representation source, embedding pooling, sample support, and gender-group coding. Without these fields, a near-zero disparity may simply reflect an undefined or inconsistently chosen target rather than a replicable fairness result.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 11, November 2026

|DOI: 10.15680/IJIRSET.2026.1511041|

H) Data Engineering and Leakage Controls

Tabular preprocessing encodes categorical attributes, imputes missing values, and preserves numeric scales only after partition construction. Learned imputers and encoders must be fitted exclusively on training observations. In COMPAS, filtering by available assessment and outcome fields also determines the observed study population; fairness indicators can change if records are excluded disproportionately by protected group. The evaluator must therefore preserve the inclusion filter and report support counts before comparing baseline and mitigated results. Label mappings are treated as part of the fitted artifact, not as a browser-session configuration.

Biography-text processing requires reproducible tokenization, truncation, embedding extraction, and class-index mapping. BERT-style pooled representations are different from GPT-2 or BLOOM token sequences, and embeddings generated by one checkpoint may be incompatible with classifiers fitted on another. The model contract includes the tokenizer revision, transformer checkpoint, pooling rule, maximum length, and occupation vocabulary. Such controls allow a result to be attributed to a specific fitted configuration and prevent misleading comparisons caused by changing feature definitions between training and inference.

I) Fairness Intervention Mechanics

Pre-processing methods intervene before classifier fitting. Reweighing modifies the contribution of records through estimated protected-group and target-label frequencies. Disparate Impact Remover repairs feature values to reduce measured dependence on sensitive membership, and Learning Fair Representations optimizes a latent encoding that balances predictive utility with fairness objectives. These methods can alter signal available for prediction; consequently, the accuracy and subgroup error response must be evaluated rather than assuming that repaired features preserve utility. Strong weighting is also unstable when a protected-label intersection contains very few observations.

In-processing and post-processing have different dependencies. Exponentiated Gradient Reduction updates a mixture of classifiers under a declared fairness constraint, which may be demographic parity or equalized odds. Reject Option Classification changes decisions within a selected uncertainty region. Equalized Odds and Calibrated Equalized Odds require group-labeled calibration data and alter the decision mechanism after predictive training. A post-processing score cannot be transferred arbitrarily to a new model checkpoint because score calibration and threshold behavior are model-dependent. Therefore, each recorded result carries the fitted intervention and the separate data partition used for its adjustment.

J) Profession Classification and Token Attribution

The profession pipeline begins with biography text and produces a fixed-dimensional representation for the selected transformer. A downstream classifier assigns scores to occupational labels. For the BERT–XGBoost variant, contextual token embeddings are pooled into document-level features and passed to boosted trees; the model cannot be interpreted as a generative language model predicting a free-form occupation string. DistilBERT, GPT-2, and BLOOM alternatives represent separate experimental configurations with their own tokenization and feature dimensions. Their measured demographic disparity is meaningful only when the downstream profession target and test records are held constant.

LIME identifies words that influence a particular classification through locally perturbed text and a surrogate prediction function. Such evidence describes the local behavior of one configured classifier. It neither estimates the causal effect of gender on the decision nor replaces dataset-level group comparisons. Repeated perturbations, unstable coefficients, truncation of occupational history, and implausible modified text can affect the explanation. The application therefore benefits from recording the predicted profession, probability vector, influential words, perturbation parameters, and model version together for reproducible inspection.

K) Protected-Group Measurement and Uncertainty

SPD compares selection rates between a declared unprivileged and privileged group. Its sign shows the direction of the outcome-rate gap, and its magnitude shows distance from demographic parity for that particular target definition. DI divides corresponding selection rates, making it sensitive to a small denominator. EOD compares true-positive rates conditional on actual favorable outcomes and requires sufficient positive-label observations in both groups. A model can have small SPD and nonzero EOD because unconditional positive rates and conditional accuracy represent different statistical properties. Such disagreements are not mathematical errors; they follow from distinct notions of fairness.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 11, November 2026

|DOI: 10.15680/IJIRSET.2026.1511041|

Fairness estimates may be unstable under rare occupational labels or sparse protected-group intersections. Before interpreting a change of a few thousandths, the audit should expose subgroup sizes, predicted-positive counts, actual-positive counts, and true-positive and false-positive components. Repeated partition evaluation or stratified bootstrap resampling can characterize variability when samples permit. Absolute SPD is particularly useful for within-task comparison when class definitions and group ordering are held fixed; it must not be treated as a universal score of equitable social outcomes.

IV. EXPERIMENTAL RESULTS

A) Reported Statistical Parity Comparison

Four signed Statistical Parity Difference values are reported for two profession-classifier configurations and two mitigation stages. Under Exponentiated Gradient Reduction (EGR), the absolute SPD is lower for BERT–XGBoost than for DistilBERT–Logistic Regression. For Calibrated Equalized Odds (CEO), BERT–XGBoost produces a negative signed value, which indicates the direction of selection-rate disparity under the chosen group order. Comparing absolute values does not establish dominance on DI, EOD, accuracy, or per-profession recall, for which matched numerical results are not provided.

TABLE I. REPORTED BIAS IN BIOS FAIRNESS RESULTS

Classifier and mitigation	SPD
DistilBERT–LR / EGR	0.009439
BERT–XGBoost / EGR	0.0008067
DistilBERT–LR / CEO	0.006604
BERT–XGBoost / CEO	−0.0040792

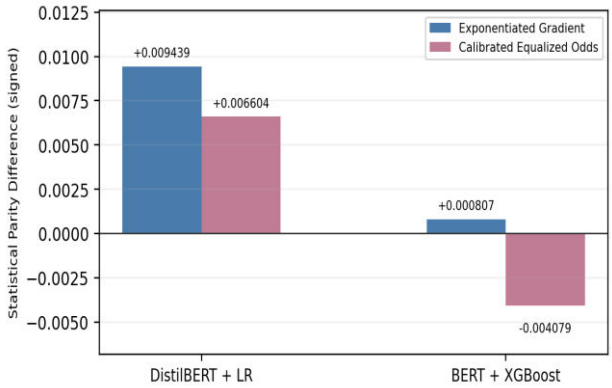


Fig.2. Signed statistical parity difference by embedding classifier and mitigation.

A small absolute SPD means that the two measured selection rates are close for the declared protected groups and audit target. It does not guarantee equal true-positive rates, false-positive rates, calibration, or representation quality. EGR and CEO are distinct mechanisms; any improvement claim must also identify subgroup denominators, favorable class, dataset partition, and the method-specific constraint.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 11, November 2026

|DOI: 10.15680/IJIRSET.2026.1511041|

B) Frontend Prediction and Audit Results

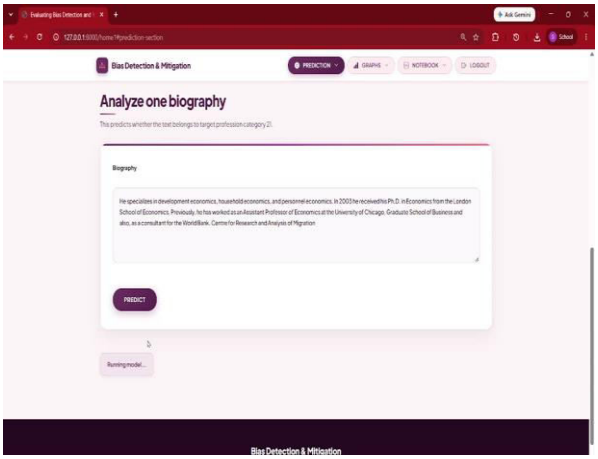


Fig. 3. Biography-input interface for transformer-based profession inference.

The biography analysis interface accepts textual occupational evidence and invokes the configured embedding-classifier pipeline. The workflow can display predicted profession and local LIME token contributions. Since LIME approximates one decision around perturbed text, its highlighted words must not be interpreted as a population-level estimate of unfairness.

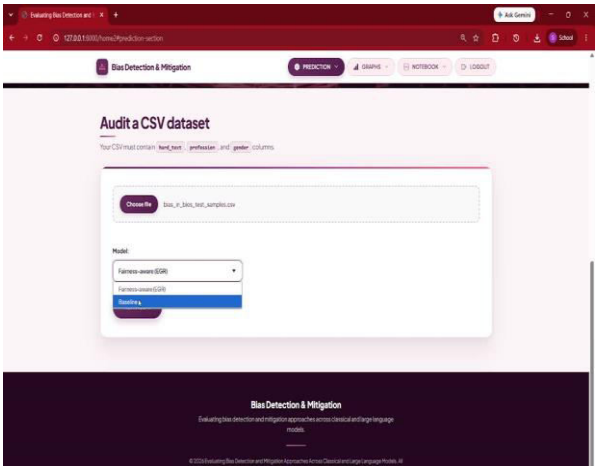


Fig. 4. Cohort-level fairness audit with CSV and model-selection controls.

The batch audit interface selects a CSV dataset and estimator configuration before group statistics are generated. Explicit biography, profession, and gender columns connect the resulting disparity estimates to the corresponding class and group contracts. Empty or unsupported groups require controlled error handling rather than silently returning zero disparity.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 11, November 2026

|DOI: 10.15680/IJIRSET.2026.1511041|

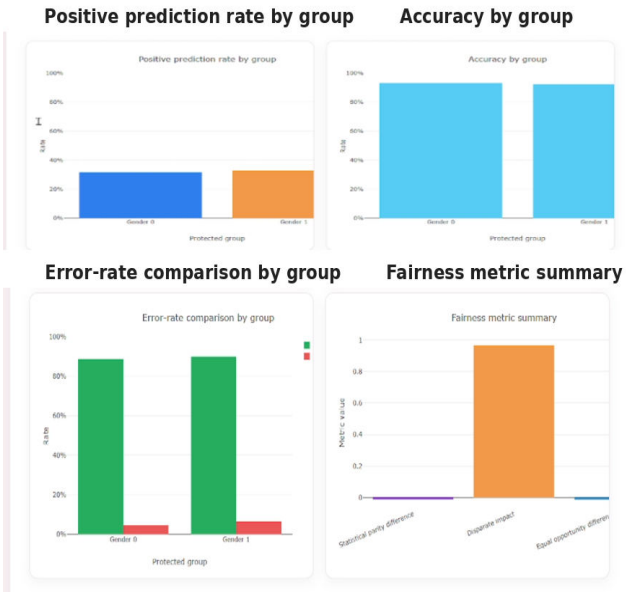


Fig. 5. Protected-group prediction rates, accuracy, error-rate comparison, and fairness statistics.

The interactive audit presents positive-prediction rates and accuracy for the two protected groups, followed by a comparison of group-specific error rates and fairness statistics. Statistical Parity Difference captures differences in favorable-decision rates, Disparate Impact expresses their ratio, and Equal Opportunity Difference compares true-positive rates conditional on the favorable class. These measures require consistent protected-group mapping, target definitions, and subgroup denominators.

C) Bias–Utility Tradeoffs

Mitigation can reduce selection-rate disparity while altering false-positive and false-negative rates. Pre-processing methods change contributions or feature distributions; constrained training changes the decision objective; post-processing changes threshold behavior after fitting. Therefore, an intervention ranking based solely on absolute SPD may select a system with reduced discrimination or worsened group-conditional errors. Model comparison should hold outcome definitions and evaluation cases constant across stages.

Occupation-frequency skew introduces an additional challenge. Gender distributions differ by profession, and one-vs-rest audits may contain very few positive observations for some occupation–group combinations. These low-support estimates can change markedly after a small number of decisions. Bootstrap intervals, category-stratified errors, and multiple training seeds would improve statistical stability. Such measurements are distinct from the recorded signed SPD comparison.

D) Reliability and Governance

The fairness calculator, explanation service, and generative narrative component have different evidence status. Numerical disparity is derived from predictions and protected-group membership; LIME offers a local approximation; and the language model produces explanatory prose. The web system should prevent generated commentary from modifying or inventing metrics, include the active model and tokenizer versions, and apply authorization and retention controls to sensitive biography content.

E) Tabular Evaluation and Sensitivity

The tabular branch serves as a methodological comparison across several fairness interventions. On Adult Income, a change in predicted high-income selection rates may arise from weight adjustment, altered feature representation, or a constrained decision rule. German Credit adds smaller-sample uncertainty and the need to distinguish good-credit and bad-credit label orientation. COMPAS adds observed recidivism-label limitations and potential differences in outcome

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 11, November 2026

|DOI: 10.15680/IJIRSET.2026.1511041|

prevalence. Each task requires the same partition and outcome definition across baseline and mitigation stages before accuracy and fairness changes can be attributed to an intervention.

A rigorous model-stage record should contain accuracy, SPD, DI, EOD, the selected constraint, and protected-group confusion counts. Aggregate accuracy without group-conditional components cannot establish whether a mitigation affects false positives or false negatives. Likewise, a reduction in unconditional selection-rate disparity cannot by itself show a reduction in disparate treatment or broader social harm. The case for a deployment candidate depends on its intended decision context, validated label meaning, model reliability, and the tradeoff selected for the affected population.

F) Model Selection and Reporting Consistency

The NLP comparison contains two mitigation stages for DistilBERT–Logistic Regression and BERT–XGBoost. EGR is a constrained learning method, while CEO adjusts calibrated classifier outputs after model fitting. This structural difference limits direct causal interpretation of the four reported SPD values. For example, the BERT–XGBoost EGR entry has a positive signed SPD of 0.0008067, whereas its CEO entry is negative at −0.0040792. Both numbers describe different locations relative to zero, but the signs must be retained so the direction of selection disparity remains visible.

The corresponding absolute disparities are 0.009439 and 0.006604 for DistilBERT–Logistic Regression and 0.0008067 and 0.0040792 for BERT–XGBoost under EGR and CEO respectively. This permits a limited ranking on selection-rate parity within the stated experiment. It does not justify an accuracy ranking or suggest that all fairness constraints have been satisfied. The absence of matched numerical DI, EOD, and full occupation-class performance rows should be handled through explicit metric scope rather than reconstructed values.

G) Interface Integrity and Audit Semantics

The browser interface distinguishes two operations that should not be conflated: single-biography inference and CSV-based cohort fairness analysis. An individual biography can yield a profession label, class scores, and local token attribution, but it does not provide a statistically meaningful estimate of demographic disparity on its own. A batch audit uses labeled records with explicit protected-group membership to calculate positive rates and conditional errors. The interface must reject missing gender, profession, or biography fields before computing or displaying group statistics. An empty subgroup should generate an undefined-metric response rather than an arbitrary zero value.

Authentication and application reliability also influence technical integrity. SQLite account records require appropriate password storage and access restrictions; uploaded biographies and CSV files may contain personally identifiable information. Saved model artifacts, tokenizers, and fairness post-processors must be versioned and loaded together. The optional Ollama explanation component should translate the deterministic audit record into prose while preserving metric signs, displayed class identities, and data support counts. If a local generation service is unavailable, a verified rule-based report prevents loss or alteration of the classifier result.

H) Threats to Validity and External Testing

Benchmark fairness findings depend on the training population, labels, protected-group definitions, and institutional setting. Adult Income reflects census-related historical conditions, German Credit represents a limited credit dataset, and COMPAS includes outcome labels formed within a justice-system observation process. Their results should not be transferred directly to a current lending or criminal-justice workflow. The profession dataset similarly contains occupational representations shaped by biography-writing conventions and occupation prevalence. Different languages, historical periods, professions, or employers may shift token associations and subgroup selection rates.

An external assessment should hold out identities and text sources when possible, vary profession-specific positive targets, and test sensitivity to tokenizer version and paragraph length. Intersectional metrics can reveal disparities concealed by single-attribute comparisons but may be unstable for small cells. Calibration and selective prediction should be checked separately from demographic parity; a classifier can be well calibrated and still exhibit group disparities. Future studies must therefore couple external prediction performance with uncertainty intervals, group supports, and independently defined fairness objectives.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 11, November 2026

|DOI: 10.15680/IJIRSET.2026.1511041|

I) Audit Reproducibility and Decision Ownership

Each fairness assessment should be associated with a complete, immutable experiment record containing the dataset release, training and evaluation partitions, sensitive-group coding, target class, model version, and mitigation stage. For post-processing methods, the record must also identify the calibration partition and selected constraint. This enables an analyst to reconstruct whether a change in SPD arose from the classifier, protected-group composition, or intervention rule. Evaluation records should preserve the positive-prediction counts underlying demographic parity and the conditional counts used in equal-opportunity analysis. Where multiple trained classifiers are compared, reporting must use the same evaluation population and favorable-class convention.

Operational responsibility must remain separate from metric optimization. A local explanation may help identify influential biography terms, while the batch audit quantifies aggregate selection and error patterns. Neither mechanism alone determines whether a specific employment, credit, or justice outcome is equitable. Human reviewers need access to the chosen fairness criterion, measured tradeoffs, uncertainty, and relevant policy context. The application can support this review by linking every visualization to the active model, audit cohort, and deterministic metric record.

J) Future Benchmark Extensions

Additional evaluation can investigate intersectional gender–occupation groups, held-out biography sources, unseen text styles, and variations in transformer pooling. Repeated partition analyses would help determine whether small reported disparity values remain stable under changes in occupational prevalence. Privacy-conscious audit design should restrict access to protected attributes and suppress publication of sparse subgroup statistics that could enable re-identification. Finally, local explanatory language-model outputs should be evaluated for consistency with metric signs, subgroup denominators, and model-specific decisions before they are shown to an analyst.

V. CONCLUSION

This work integrates classical classifier fairness analysis and transformer-based occupation prediction within a traceable evaluation and web-audit framework. Adult Income, German Credit, COMPAS, and Bias in Bios provide distinct targets and protected-attribute contracts. Reweighing, feature repair, constrained learning, and post-processing methods support comparisons between predictive utility and fairness. The recorded Bias in Bios results identify BERT–XGBoost with Exponentiated Gradient as having the lowest absolute SPD among the four reported configurations. Conclusions about broader equity require matched accuracy, DI, EOD, group support, and external-validation evidence. Future development should strengthen subgroup uncertainty measurement, LIME stability, data governance, and consistency checks for generated explanations.

REFERENCES

[1] Gallegos, I. O., et al. (2024). Bias and Fairness in Large Language Models: A Survey. Computational Linguistics, 50(3), 1097–1179. https://aclanthology.org/2024.cl-3.8/

[2] Zhong, M., and Tandon, R. (2024). Intrinsic Fairness-Accuracy Tradeoffs under Equalized Odds. https://arxiv.org/abs/2405.07393

[3] Dablain, D., Krawczyk, B., and Chawla, N. V. (2024). Towards a holistic view of bias in machine learning: bridging algorithmic fairness and imbalanced learning. Discover Data. https://link.springer.com/article/10.1007/s44248-024-00007-1

[4] Liu, Z., et al. (2024). AIM: Attributing, Interpreting, Mitigating Data Unfairness. KDD 2024. https://research.ibm.com/publications/aim-attributing-interpreting-mitigating-data-unfairness

[5] Fawkes, J., Fishman, N., Andrews, M., and Lipton, Z. C. (2024). The Fragility of Fairness: Causal Sensitivity Analysis for Fair Machine Learning. NeurIPS 2024. https://papers.nips.cc/paper_files/paper/2024/hash/f7be3ebca4980b59fe3f665011115395-Abstract-Datasets_and_Benchmarks_Track.html

[6] Adewumi, T., et al. (2024). Fairness and Bias in Multimodal AI: A Survey. https://arxiv.org/abs/2406.19097

[7] Hoitsma, F., et al. (2024, online). Mitigating implicit and explicit bias in structured data without sacrificing accuracy in pattern classification. AI & Society. https://link.springer.com/article/10.1007/s00146-024-02003-0

[8] Bellamy, R. K. E., et al. (2019). AI Fairness 360: An Extensible Toolkit for Detecting and Mitigating Algorithmic Bias. IBM Journal of Research and Development. https://arxiv.org/abs/1810.01943

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 11, November 2026

|DOI: 10.15680/IJIRSET.2026.1511041|

[9]Agarwal, A., et al. (2018). A Reductions Approach to Fair Classification. ICML. https://proceedings.mlr.press/v80/agarwal18a.html
[10] Hardt, M., Price, E., and Srebro, N. (2016). Equality of Opportunity in Supervised Learning. NeurIPS. https://arxiv.org/abs/1610.02413
[11] Pleiss, G., et al. (2017). On Fairness and Calibration. NeurIPS. https://arxiv.org/abs/1709.02012
[12] Kamiran, F., and Calders, T. (2012). Data Preprocessing Techniques for Classification without Discrimination. Knowledge and Information Systems. https://doi.org/10.1007/s10115-011-0463-8
[13] Zemel, R., et al. (2013). Learning Fair Representations. ICML. https://proceedings.mlr.press/v28/zemel13.html
[14] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Why Should I Trust You? Explaining the Predictions of Any Classifier. KDD. https://arxiv.org/abs/1602.04938
[15]Chen, T., and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. KDD. https://arxiv.org/abs/1603.02754
[16] Devlin, J., et al. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL. https://aclanthology.org/N19-1423/
[17] Sanh, V., et al. (2019). DistilBERT, a distilled version of BERT. https://arxiv.org/abs/1910.01108
[18] Radford, A., et al. (2019). Language Models are Unsupervised Multitask Learners. OpenAI Technical Report. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
[19]Scao, T. L., et al. (2022). BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. https://arxiv.org/abs/2211.05100
[20] De-Arteaga, M., et al. (2019). Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting. FAT*. https://arxiv.org/abs/1901.09451
[21]Angwin, J., et al. (2016). Machine Bias. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
[22]Becker, B., and Kohavi, R. (1996). Adult. UCI Machine Learning Repository. https://archive.ics.uci.edu/dataset/2/adult
[23]Hofmann, H. (1994). Statlog (German Credit Data). UCI Machine Learning Repository. https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data
[24] ProPublica. (2016). COMPAS Analysis Data and Code. https://github.com/propublica/compas-analysis
[25] Fairlearn Contributors. Fairlearn documentation: Assessment and mitigation. https://fairlearn.org/
[26] AIF360 Contributors. AI Fairness 360 Documentation. https://aif360.readthedocs.io/
[27] FastAPI Project. FastAPI Documentation. https://fastapi.tiangolo.com/
[28] Hugging Face. Transformers Documentation. https://huggingface.co/docs/transformers/
[29] Ollama. Ollama Documentation. https://docs.ollama.com/
[30]Anaconda. Anaconda Distribution Documentation. https://www.anaconda.com/docs/getting-started/anaconda/install



INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details