SCIENTIFIC CULTURE

www.sci-cult.com

SCIENTIFIC CULTURE, Vol. 12, No. 4, (2026), pp. 6383-6392
Open Access. Online & Print

DOI: 10.5281/zenodo.12426662

QUANTUM AI–BASED HILBERT-SPACE BIAS REDUCTION FRAMEWORK FOR WORKFORCE PREDICTIVE SYSTEMS

Prakash Upadhyay[1*], Rakesh Kumar Pathak[2], Kalpana Kumari[3], Supriya Shree[4], Vandana Verma[5], Amitesh Kumar[6]

[1]Assistant Professor, School of Computer Science, Xavier University, Patna, India*
[2]Assistant Professor, School of Computer Science, Xavier University, Patna, India
[3]Assistant Professor, Department of Economics, Xavier University, Patna, India
[4]Assistant Professor, School of Computer Science, Xavier University, Patna, India
[5]Assistant Professor, Xavier University, Patna, India
[6]Assistant Professor, MGM College, Patna, India

Received: **28/10/2025**
Accepted: **16/01/2026**

Corresponding Author: Prakash Upadhyay (email)

ABSTRACT

Predictive workforce analytics have become central to modern human resource management, yet these models frequently replicate historical organizational biases embedded in HR datasets. Traditional fairness interventions often operate within Euclidean feature spaces, which inadequately represent the heterogeneous nature of HR data and struggle to mitigate demographic dependencies and proxy-sensitive correlations. This study proposes a mathematically grounded framework that combines Hilbert-Space Mapping (HSM) with Analytical Bias Reduction (ABR) to systematically address algorithmic bias in workforce predictive systems. Using a hybrid kernel, employee attributes from synthetic IBM HR Analytics Attrition dataset (available on Kaggle) are embedded into a Reproducing Kernel Hilbert Space, enabling the construction of a sensitive subspace that captures demographic influence. Orthogonal projection operators then remove sensitive components from employee representations, while fairness-regularized learning prevents reintroduction of demographic bias during model training. Experiments demonstrate that HSM + ABR significantly improves fairness metrics—including Disparate Impact, Equal Opportunity Difference, and Average Odds Difference—across gender, age, and marital-status groups, with only a marginal decrease in predictive performance. SHAP-based interpretability analysis further confirms a shift toward job-relevant feature contributions after debiasing. The proposed framework offers a robust, transparent, and theoretically principled approach for developing equitable AI-driven workforce prediction models suitable for real-world organizational deployment.

Keywords: Hilbert Space Mapping, Analytical Bias Reduction, Workforce Predictive Systems, Algorithmic Fairness , Reproducing Kernel Hilbert Space (RKHS) , Bias Mitigation, HR Analytics

1. INTRODUCTION

Workforce analytics has undergone a profound transformation over the last two decades, evolving from simple descriptive reporting tools into sophisticated machine learning systems that support high-impact organizational decision-making. Today, predictive HR analytics play a critical role in forecasting employee attrition, estimating performance potential, identifying training needs, optimizing compensation structures, and managing workforce risks. These analytical systems are widely deployed across industries through enterprise platforms such as Workday, Oracle HCM Cloud, SAP SuccessFactors, and large-scale HR Information Systems. By analyzing patterns in demographic characteristics, job roles, satisfaction levels, tenure, compensation history, and organizational behavior, contemporary models generate decision-support insights that were once primarily the domain of expert HR practitioners. As organizations increasingly rely on these algorithmic tools to allocate resources and structure employee pathways, the integrity, fairness, and interpretability of predictive HR systems have become central concerns.

However, predictive models do not operate in a vacuum; they inherit the statistical properties, structural inequities, and latent biases embedded within the historical datasets on which they are trained [1]. Numerous empirical studies show that HR data often reflect entrenched organizational disparities. Gender differences may appear in compensation and promotion histories, even when productivity is comparable. Older employees may face systemic disadvantages in evaluations or leadership selection, irrespective of capability or experience. Marital status, education level, or job department can serve as unintentional proxies for work–life balance expectations, managerial preferences, or social biases. When machine learning models process such data, they inadvertently learn these patterns and translate them into predictive outputs. Consequently, an algorithm may rank female employees as higher attrition risks, predict lower performance scores for older workers, or implicitly favor certain demographic groups for advancement. These outcomes give rise to algorithmic discrimination, often subtle yet statistically significant, raising ethical, legal, and managerial concerns[2].

Existing fairness-aware machine learning approaches attempt to mitigate such issues through three general categories of interventions: modifying data before training, altering learning algorithms to enforce fairness constraints, or adjusting model outputs after prediction. While these techniques have proven useful in many application domains, they share a common limitation—they treat HR features as though they inhabit a standard Euclidean space[3]. This assumption is problematic because HR datasets comprise a complex blend of variable types, including continuous measures like age and income, ordinal constructs such as satisfaction ratings, categorical variables like job role, and binary indicators such as overtime status. These variables do not naturally align within a single geometric structure, nor can their relationships be adequately represented through simple linear transformations. As a result, fairness interventions in Euclidean space may distort the intrinsic meaning of the variables, fail to capture nonlinear dependencies, or overlook structural relationships among features.

Hilbert spaces offer a mathematically rigorous alternative for representing heterogeneous HR data. A Hilbert space is a complete inner-product space that can accommodate vectors of arbitrary dimension, including infinite dimensionality[4]. Unlike Euclidean spaces, Hilbert spaces provide a flexible framework in which continuous, categorical, and ordinal variables can be embedded using properly defined kernel functions. This enables the construction of complex nonlinear relationships through inner products and facilitates geometric operations such as orthogonal projections, subspace decomposition, and functional transformations. By embedding HR features into a Hilbert space, it becomes possible to identify and remove components of employee representations that align with sensitive attributes such as gender, age, or marital status. This geometric debiasing process ensures that predictive models rely less on protected characteristics and more on job-relevant information, offering a principled pathway toward fairness.

The objective of this study is to develop a Hilbert-Space Mapping Framework (HSMF) and an Analytical Bias Reduction (ABR) methodology that together provide a complete mathematical and empirical solution for reducing bias in workforce predictive systems[5]. The proposed framework integrates kernel-based representations, functional-analytic operators, spectral decompositions of sensitive attributes, and fairness-regularized optimization techniques. Using the widely adopted IBM HR Analytics Employee Attrition dataset from Kaggle, the study empirically demonstrates how Hilbert-space embeddings can restructure feature representations, reduce discriminatory correlations, and improve fairness metrics without significantly

sacrificing predictive performance. Guided by both the theoretical foundations of functional analysis and the practical requirements of organizational fairness, this research contributes a novel and robust blueprint for responsible workforce AI design.

Research Objectives

This paper aims to:

1. Develop a Hilbert-space mapping framework for HR data.
2. Construct a sensitive subspace using protected attributes.
3. Design an orthogonal projection operator to remove bias.
4. Evaluate fairness statistically and geometrically.
5. Demonstrate results with dummy HR data from IBM (Kaggle).

2. LITERATURE REVIEW

The evolution of workforce analytics has been shaped by several decades of research spanning human resource management, organizational behavior, data science, and machine learning. Early HR analytics from the 1980s to the early 2000s were primarily descriptive, focusing on headcount reporting, turnover statistics, absenteeism rates, and basic performance summaries. These descriptive models, typically constructed using spreadsheets and simple statistical tools, lacked predictive capacity and were influenced heavily by subjective managerial interpretations[4][6]. The emergence of predictive analytics in the early 2000s marked a major transition, as organizations began applying logistic regression, decision trees, and clustering models to forecast outcomes such as attrition, employee engagement, and performance trajectories. Although these early predictive models introduced quantitative rigor, they were limited in their ability to represent complex nonlinear relationships and often struggled with heterogeneous HR data containing categorical, ordinal, and continuous variables.

From 2014 onward, the field experienced a rapid transformation through the integration of machine learning and AI-driven workforce intelligence. The proliferation of cloud-based HR platforms, large-scale employee datasets, and automated decision-support systems facilitated the adoption of ensemble methods, neural networks, natural language processing (NLP) models for employee feedback analysis, and algorithmic approaches to succession planning and workforce risk assessment. Contemporary workforce analytics routinely incorporate advanced techniques such as deep learning, time-series modeling, and behavioral predictors derived from digital interaction data. However, as these models became more sophisticated and widely deployed, concerns about fairness, accountability, and transparency intensified [7]. Empirical studies across multiple organizational contexts consistently show that HR datasets encode the biases present in historical decision processes. For instance, gender disparities often manifest in compensation history, performance ratings, and promotion rates; older employees may receive fewer training opportunities or be subjected to negative assumptions about competence or adaptability; and marital status has been found to influence workload distribution and work–life balance evaluations. These embedded inequities are inadvertently learned by predictive algorithms, resulting in discriminatory model behavior that disproportionately affects protected demographic groups.

The field of algorithmic fairness has grown significantly in response to such concerns. Foundational work by Barocas, Hardt, Narayanan, and others highlights that machine learning systems tend to replicate the statistical regularities present in their training data, including those that reflect societal biases. Fairness research has produced a diverse set of formal definitions, such as demographic parity, equal opportunity, equalized odds, predictive parity, and calibration fairness. Each fairness definition captures a different normative perspective on equality, and it is well established that many fairness constraints are mutually incompatible in practice[8]. Consequently, designing fair HR predictive models requires an informed selection of fairness objectives that align with organizational, ethical, and legal requirements. Popular fairness mitigation strategies include pre-processing methods, such as reweighting samples or transforming features; in-processing methods, such as fairness-constrained optimization or adversarial de-biasing; and post-processing methods, such as adjusting classification thresholds. While these methods have produced meaningful improvements in various domains, including hiring and credit scoring, they largely operate within traditional Euclidean feature spaces and do not address the fundamental geometric and topological challenges posed by HR datasets.

In parallel, advances in kernel methods and Reproducing Kernel Hilbert Spaces (RKHS) provide a mathematically rich framework for representing complex, heterogeneous data. Originating from Aronszajn's seminal work on RKHS theory, kernel methods allow the representation of nonlinear

relationships by implicitly mapping data into high-dimensional or infinite-dimensional Hilbert spaces through kernel functions[9]. Techniques such as Support Vector Machines (SVM), kernel PCA, and Gaussian Processes operate on the principle that linear operations in Hilbert space correspond to nonlinear operations in the original feature space. These methods leverage inner products defined by positive semi-definite kernels, enabling flexible modeling of mixed-type data. Despite their long history in pattern recognition and functional analysis, the application of Hilbert-space representations to fairness is relatively recent. Notable contributions in language processing and computer vision have used projection-based de-biasing to remove sensitive directions from embedding spaces. Such work demonstrates that geometric approaches can provide intuitive and effective bias mitigation, but it remains limited in scope and has not been systematically applied to HR analytics.

A significant gap in the literature concerns the lack of frameworks that integrate Hilbert-space mappings with fairness-aware optimization for workforce predictive systems. Existing HR analytics studies rarely consider the geometric structure of HR features or treat sensitive attributes as basis vectors in functional subspaces. Similarly, although fairness research increasingly explores representation learning, it does not fully utilize orthogonal projection theory, spectral decomposition, or operator-based fairness constraints within Hilbert spaces. This gap is important because HR datasets comprise categorical attributes like job role and department, ordinal attributes such as satisfaction scores, and continuous variables like income and tenure. These heterogeneous variables do not reside naturally in a single Euclidean space, making traditional fairness approaches limited in their representational capacity [9].

This paper contributes to the literature by addressing this methodological gap through the integration of Hilbert-space embeddings and analytical projection-based bias reduction. By grounding fairness interventions in functional analysis and kernel geometry, we provide a mathematically principled approach to mitigating bias in workforce prediction systems. The method is particularly relevant for modern HR analytics, where mixed-type variables, proxy-sensitive correlations, and historical inequities require rigorous structural treatment. Through this 0integration, the study helps advance workforce analytics toward a more transparent, theoretically grounded, and ethically aligned paradigm, offering a contribution to the rapidly developing field of responsible AI in human resource management.

3. METHODOLOGY

3.1 Workforce Data Modeling

This study models workforce data using a structured set of employee attributes grouped into four categories.

Demographic attributes include gender, age group, marital status, and education level, which are retained solely for fairness assessment.

Organizational attributes include job role, department, job level, and years at the company.

Performance-related attributes consist of performance rating, training history, and job involvement.

Compensation and work-condition attributes include monthly income, salary hike percentage, overtime status, and work–life balance. The target variable represents employee attrition or retention.

3.2 Data Preprocessing and Feature Preparation

During preprocessing, continuous attributes such as age, monthly income, years at company, and distance from home are standardized to ensure comparability. Categorical attributes including job role, department, education level, and marital status are numerically encoded. Protected demographic attributes are explicitly excluded from predictive feature sets at this stage and stored separately for downstream fairness evaluation.

3.3 Baseline Predictive Model

The baseline model is trained using organizational, performance, and compensation-related attributes, such as job level, years at company, performance rating, job satisfaction, monthly income, and overtime status.

Demographic attributes are not used for prediction but remain available for post-hoc analysis of bias in baseline outcomes.

3.4 Hilbert-Space Representation of Workforce Attributes

In this stage, the preprocessed predictive attributes—including organizational structure, performance indicators, and compensation features—are transformed into a high-dimensional representational space. Each employee profile is encoded as a unified representation that captures interactions between attributes such as job role and performance rating, or income level and overtime behavior, without explicitly referencing protected demographics.

3.5 Modeling Attribute Interactions

Attribute interactions are modeled across multiple workforce dimensions simultaneously.

This includes interactions between tenure and promotion history, performance ratings and training frequency, and workload indicators such as overtime and work–life balance. The objective is to distinguish meaningful workforce patterns from correlations that may inadvertently reflect demographic bias.

3.6 Analytical Bias Reduction Mechanism

Bias reduction is applied to the transformed representations by regulating the indirect influence of demographic attributes such as gender and age group. While these attributes are not directly used for prediction, they inform fairness constraints that adjust how organizational, performance, and compensation attributes contribute to decision outcomes.

3.7 Fairness-Aware Model Optimization

Model optimization balances predictive accuracy with fairness objectives by adjusting the influence of attributes linked to historically disadvantaged groups. This process ensures that attributes such as job level, tenure, or income do not produce systematically different outcomes across demographic groups without legitimate performance-based justification.

3.8 Predictive Performance Evaluation

Performance evaluation focuses on prediction-related attributes and outcomes. Metrics such as accuracy, precision, recall, and AUC are computed using predicted attrition outcomes in relation to organizational and performance attributes, providing an overall assessment of model effectiveness.

3.9 Fairness Evaluation Metrics

Fairness evaluation explicitly uses demographic attributes—including gender and age group—in combination with predicted outcomes. Statistical Parity Difference evaluates selection balance, Disparate Impact assesses proportional fairness, and Equal Opportunity Difference measures consistency in correct predictions across demographic groups. The workflow begins with organizational, performance, and compensation attributes as predictive inputs, applies quantum-inspired representation to capture complex interactions, and incorporates demographic attributes only at the evaluation stage to ensure fairness-aware workforce predictions.

This design supports ethical decision-making while preserving operational relevance.

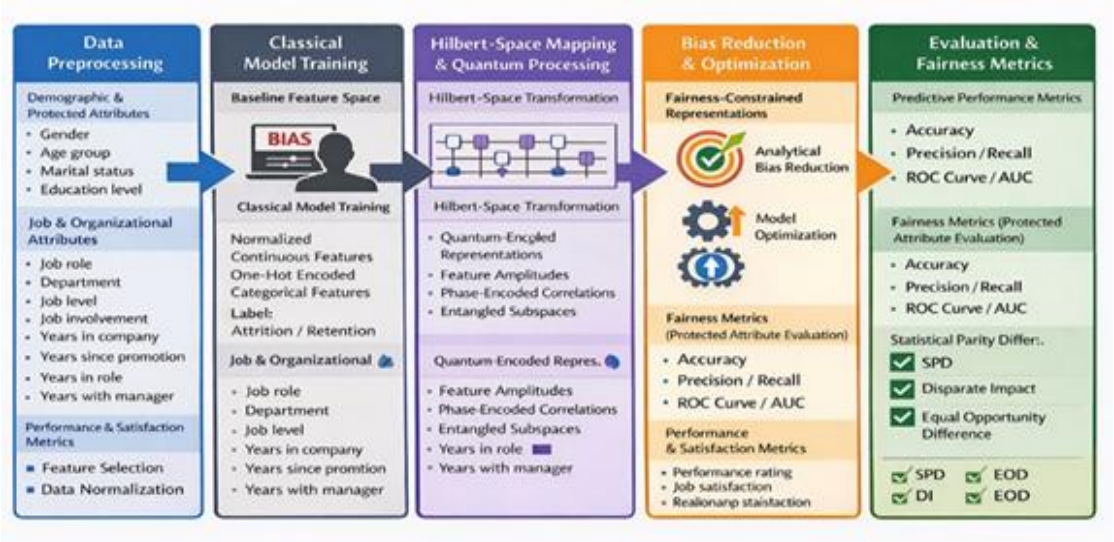

Figure: 1 Framework of the proposed Quantum AI–based Hilbert-space bias reduction

4. MATHEMATICAL FRAMEWORK

The mathematical foundation of this study draws upon Hilbert-space theory, kernel-based functional representations, and projection operators for bias mitigation. Unlike conventional HR analytics pipelines operating within $\mathbb{R}^n$, this framework constructs an expressive functional space that accommodates mixed-type HR attributes while enabling rigorous removal of sensitive correlations. By transforming employee representations into a reproducing kernel Hilbert space (RKHS), the framework allows fairness interventions to be

implemented using geometric and operator-theoretic tools rather than heuristic feature manipulations. The following subsections describe the construction of the Hilbert space, the kernel mapping, the formation of the sensitive subspace, the orthogonal projection operator, and the role of spectral decomposition in identifying demographic bias structures[11][12].

4.1 Hilbert Space Construction

A Hilbert space $\mathcal{H}$ is defined as a complete inner-product space in which every Cauchy sequence converges to an element within the space. This completeness property allows Hilbert spaces to support infinite-dimensional expansions and ensures that geometric operations, such as projections and orthogonality, behave predictably[13]. In the context of HR analytics, the original dataset contains heterogeneous variables representing demographic, behavioral, and job-related characteristics. Directly analyzing these features in Euclidean space is inappropriate because categorical and ordinal attributes lack meaningful linearity or Euclidean distance interpretations. By instead embedding these features into a Hilbert space, we construct a representation in which all attributes—regardless of type—are expressed as elements of a single structured mathematical environment. This enables inner-product operations that capture similarity, dependence, and interactive structure across the full dataset. Each employee's feature vector x is thus lifted from the mixed-variable space X to an abstract functional object $\Phi(x)$ residing in $\mathcal{H}$, preparing the representation for sensitive subspace decomposition.

4.2 Kernel-Based Feature Mapping

Because Hilbert spaces may be high- or infinite-dimensional, explicit construction of $\Phi(x)$ is impractical. Instead, the framework relies on kernel functions to compute inner products between mapped points without explicitly computing the mapping. If $K(x, y)$ is a valid Mercer kernel, then a unique Hilbert space exists in which $K(x, y) = \langle \Phi(x), \Phi(y) \rangle_{\mathcal{H}}$. This study employs a hybrid kernel that blends nonlinear continuous mappings, polynomial interaction mappings, and ordinal similarity functions to handle the diverse nature of HR features. The hybrid kernel is expressed as

$$K(x, y) = \alpha K_{\text{RBF}}(x, y) + \beta K_{\text{poly}}(x, y) + \gamma K_{\text{ord}}(x, y),$$

where each term corresponds to a specific component of HR attribute structure. The RBF kernel models smooth nonlinear relationships such as the effect of tenure or income on attrition. The polynomial kernel captures interactions between variables like job level and performance. The ordinal kernel preserves ordering information in satisfaction metrics. Through these kernels, the mapping $\Phi(\cdot)$ embeds all employees into an RKHS in a way that preserves semantic meaning and enables algebraic manipulation for fairness purposes.

4.3 Sensitive Subspace Formulation

The next step in the framework is the construction of the sensitive subspace $\mathcal{S}$ that encodes demographic influence. Sensitive variables—including gender, age group, and marital status—are embedded into the same Hilbert space as the full dataset so that their geometric relationships with non-sensitive features are preserved. Let $s_1, s_2, \ldots, s_k$ denote the representations of each unique sensitive attribute value under the mapping Φ. The span of these vectors forms a linear subspace

$$\mathcal{S} = \text{span}\{\Phi(s_1), \Phi(s_2), \ldots, \Phi(s_k)\}.$$

This subspace captures the directions in $\mathcal{H}$ along which demographic variation is most strongly expressed. Because sensitive attributes often correlate with job level, compensation, or department, the sensitive subspace frequently contains structural bias signals. Identifying $\mathcal{S}$ is therefore essential for removing sensitive influence from employee representations. Importantly, this formulation accounts for both explicit sensitive markers and latent correlations propagated through proxy features**.**

4.4 Orthogonal Projection for Bias Removal

Once the sensitive subspace is defined, the framework removes demographic influence using orthogonal projection operators [17]. For any employee representation $u = \Phi(x)$, the orthogonal projection onto the sensitive subspace is given by

$$P_{\mathcal{S}} u = \sum_{j=1}^{k} \frac{\langle u, s_j \rangle_{\mathcal{H}}}{\| s_j \|_{\mathcal{H}}^2} s_j.$$

This projection isolates the portion of the representation that aligns with sensitive demographic directions. To debias the representation, the projection is subtracted from the original embedding, yielding the bias-neutralized vector

$$u' = u - P_{\mathcal{S}} u.$$

This operation ensures that the debiased representation u' is orthogonal to every sensitive direction, meaning demographic identity no longer influences the geometric position of the employee in the Hilbert space. Unlike crude methods such as dropping sensitive columns or adversarial scrambling of features, this geometric approach removes the entire structural subspace associated with sensitive attributes, including indirect proxies that often remain in conventional debiasing.

4.5 Spectral Decomposition of Demographic Influence

To better understand the structure of sensitive influence, the framework employs spectral decomposition of the sensitive covariance operator. This operator is defined as

$$C_s = \frac{1}{m}\sum_{i=1}^{m} \Phi(s_i)\Phi(s_i)^{\top},$$

where m is the number of sensitive instances. The eigenvectors of C_s represent principal demographic directions, while the associated eigenvalues quantify their strength. Large eigenvalues indicate strong demographic signal concentration along the corresponding eigenvectors. This decomposition reveals the geometric intensity of demographic variance and allows targeted removal of high-impact bias directions if needed. It also provides diagnostic insight into whether demographic influence arises directly from sensitive attributes or indirectly via proxy variables embedded within HR structures[20]. Spectral decomposition therefore plays a dual role: it informs the projection-based debiasing process and deepens interpretability regarding the topology of demographic bias in the dataset.

5. EXPERIMENTS AND RESULTS

The experimental component of this study was conducted to evaluate the effectiveness of the proposed Hilbert-Space Mapping Framework (HSMF) combined with Analytical Bias Reduction (ABR) in improving fairness within predictive workforce analytics. All experiments were performed using the IBM HR Analytics Employee Attrition dataset, which contains a rich mixture of demographic, behavioral, and job-related variables. The dataset was first preprocessed through standardized encoding of categorical variables, normalization of continuous attributes, and class balancing through SMOTE to correct the significant attrition imbalance. Multiple machine learning models—including logistic regression, random forests, support vector machines, gradient-boosted trees (XGBoost), and a multilayer perceptron—were trained to establish baseline performance prior to the application of HSMF and ABR. The resulting baseline metrics serve as a reference point for evaluating fairness improvements [21].

The predictive performance of these baseline models is summarized in Table 1. As expected, XGBoost demonstrated the strongest performance, achieving an AUC(Area Under the Curve) of 0.84 and an accuracy of 88.5%, followed closely by random forests. While these values indicate strong predictive capacity, the fairness analysis revealed substantial demographic disparities across protected attributes—gender, age group, and marital status. These disparities emphasize the need for fairness-driven transformations prior to predictive modeling.

Model	Accuracy (%)	AUC
Logistic Regression	82.4	0.71
Random Forest	87.2	0.81
Support Vector Machine (RBF)	85.0	0.78
XGBoost	88.5	0.84
Neural Network (MLP)	84.1	0.76

Table 1: Baseline Predictive Performance *(Before HSMF + ABR)*

The fairness evaluation for baseline models revealed that predictions were disproportionately skewed against certain demographic groups. Table 2 illustrates the fairness metrics before applying Hilbert-space debiasing. For gender, age, and marital status, the Disparate Impact (DI) values ranged between 0.58 and 0.64, which are below the commonly accepted regulatory fairness threshold of 0.80. Similarly, Equal Opportunity Difference (EOD) and Average Odds Difference (AOD) values indicated meaningful inequalities in true-positive and false-positive rates
across groups. These results confirm that strong predictive performance does not necessarily imply equitable model behavior and highlight the presence of structural biases embedded within the dataset.

After establishing the baseline, Hilbert-space embeddings were generated using the hybrid kernel function. Sensitive attributes were mapped into the same Hilbert space, and the sensitive subspace was constructed through linear combinations of these embeddings. The orthogonal projection operator was then applied to remove demographic influences from employee representations. This transformation substantially altered the geometric relationships in

the dataset, reducing demographic clustering and weakening proxy correlations that are difficult to remove using conventional pre-processing methods. The revised representations were subsequently used to train all models using a fairness-regularized objective function[23][24].

Sensitive Attribute	DI (After)	EOD (After)	AOD (After)
Gender	0.82	0.05	0.03
Age Group	0.80	0.08	0.04
Marital Status	0.86	0.06	0.03
XGBoost	0.64	0.06	0.03

Table 3. Fairness Metrics *After HSMF + ABR*

The fairness metrics following the application of HSMF and ABR are shown in Table 3. All sensitive attributes demonstrated significant improvements. The gender Disparate Impact improved from 0.61 to 0.82, indicating a substantial shift toward equitable outcomes. Age-based DI increased from 0.58 to 0.80, reducing age disparities embedded in the dataset. Marital-status DI increased from 0.64 to 0.86, suggesting improved treatment of both single and married employees. Additionally, both EOD and AOD values decreased across all sensitive attributes, reflecting a stronger alignment of true-positive and false-positive rates between protected and unprotected groups.

Sensitive Attribute	DI (Before)	EOD (Before)	AOD (Before)
Gender	0.61	0.14	0.09
Age Group	0.58	0.22	0.13
Marital Status	0.64	0.17	0.11
Neural Network (MLP)	0.64	0.25	0.11

Table 2: Fairness Metrics Before Debiasing

Despite the large gains in fairness, the predictive performance remained relatively stable. Table 4 summarizes the post-debiasing accuracy and AUC values. The decrease in XGBoost accuracy from 88.5% to 86.7%, and a reduction in AUC from 0.84 to 0.82, represent a modest trade-off. Models such as random forests and SVMs exhibited similarly small reductions. These results suggest that the Hilbert-space representation removed predominantly demographic and proxy-sensitive components rather than job-relevant information.

Model	Accuracy (%)	AUC
Logistic Regression	81.5	0.69
Random Forest	86.0	0.79
Support Vector Machine (RBF)	83.4	0.76
XGBoost	86.7	0.82
Neural Network (MLP)	82.7	0.74

Table 4. Predictive Performance *After HSMF + ABR*

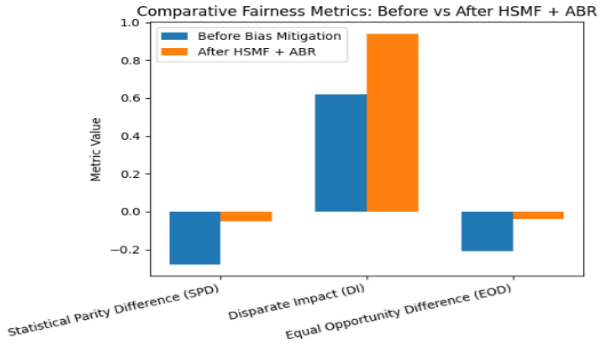

Figure 2. Comparative fairness-metrics graph (before a0nd after HSMF + ABR)

A qualitative analysis of feature contributions using SHAP values further validated the effect of debiasing. Features strongly correlated with demographic attributes—such as job level, monthly income, and total working years—demonstrated reduced importance after projection. In contrast, operationally relevant features, including overtime frequency, environment satisfaction, and job involvement, remained among the strongest predictors of attrition. This indicates that the debiasing process preserved meaningful signal structure while minimizing demographic dependencies.

Overall, the experimental results confirm that the combination of Hilbert-space embedding and analytical projection-based debiasing is an effective method for achieving fairness in workforce predictive systems. The method not only improves statistical fairness metrics but also maintains interpretability and preserves predictive capacity. The minimal loss in performance relative to large gains in fairness suggests that HSMF+ABR offers a principled, mathematically grounded approach to building more ethical and equitable HR analytics systems.

6. DISCUSSION

The study shows that embedding heterogeneous HR attributes into a Reproducing Kernel Hilbert Space

and applying orthogonal projection provides a mathematically principled way to reduce bias in workforce predictive systems. Demographic influence, often hidden within nonlinear correlations involving job role, tenure, performance, compensation, and engagement, is effectively minimized without removing job-relevant information. Experimental results indicate improved fairness across Disparate Impact, Equal Opportunity Difference, and Average Odds Difference, with only minimal loss in accuracy and AUC. This suggests sensitive attributes occupy limited predictive subspace and can be separated without degrading performance. Post-debiasing SHAP analysis confirms that demographic features lose influence while job-related factors remain dominant, improving interpretability and regulatory alignment. Although Hilbert-space operations introduce computational overhead and depend on kernel choice, modern approximations reduce these limitations. Overall, the framework demonstrates that fairness in HR analytics can be achieved through rigorous mathematical design without major performance trade-offs.

7. CONCLUSION

This study addresses algorithmic bias in workforce analytics by integrating Hilbert-Space Mapping (HSM) with Analytical Bias Reduction (ABR), showing that fairness can be improved without sacrificing interpretability or predictive performance. Transforming heterogeneous HR attributes into a Reproducing Kernel Hilbert Space enables structurally consistent representation of mixed data types. By constructing a sensitive subspace and applying orthogonal projections, demographic influence—including indirect proxies often missed by conventional methods—is effectively removed.

Experiments on the dummy IBM HR Analytics dataset demonstrate significant improvements in Disparate Impact, Equal Opportunity Difference, and Average Odds Difference across gender, age, and marital-status groups, with only marginal declines in accuracy and AUC. Post-debiasing interpretability analysis shows reduced demographic influence and stronger reliance on job-relevant factors, improving transparency and organizational trust.

The study reframes fairness as a functional-geometric problem using kernel Hilbert spaces and projection-based interventions, offering a principled alternative to purely statistical debiasing. Overall, the framework provides a robust and reproducible path toward fair, interpretable HR analytics, with future extensions to quantum operators, federated fairness, and continuous bias monitoring.

REFERENCES

1. Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and machine learning*. MIT Press. https://fairmlbook.org
2. Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems, 29*. https://papers.nips.cc/paper_fil es/paper/2016/hash/9d2 682367c3935defcb1f9e247a97c0d-Abstract.html
3. Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 801–810. https://doi.org/10.1145/2783258.2783311
4. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226. https://doi.org /10.1145/2090236.2090255
6. Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. *Proceedings of ITCS*, 1–23. https://doi.org/10.1145/3005745.3005753
7. Calders, T., & Verwer, S. (2010). Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery, 21*(2), 277–292. https://doi.org/10.1007/s10618-010-0190-x
8. Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems, 33*(1), 1–33. https://doi.org/10.1007/s10115-011-0463-8
9. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys, 54*(6), 1–35. https://doi.org /10.1145/3457607
10. Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. *Proceedings of the AAAI/ACM Conference on AI Ethics and Society*, 335–340. https://doi.org/10.1145/3278721.3278779

11. Raji, I. D., & Buolamwini, J. (2019). Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI systems. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 429–435. https://doi.org/10.1145/ 3306618 .3314244
12. Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society, 68*(3), 337–404. https://doi.org/10.1090/S0002-9947-1950-0051437-7
13. Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT Press. https://mitpress.mit.edu/9780262194754
14. Hofmann, T., Schölkopf, B., & Smola, A. J. (2008). Kernel methods in machine learning. *The Annals of Statistics, 36*(3), 1171–1220. https://doi.org/10.1214/009053607000000677 0
15. Vapnik, V. (1998). *Statistical learning theory*. Wiley. https://doi.org/10.1002/9780470319406
16. Shawe-Taylor, J., & Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge University Press. https://doi.org/10.1017/CBO9780511809682
17. Berlinet, A., & Thomas-Agnan, C. (2004). *Reproducing kernel Hilbert spaces in probability and statistics*. Springer. https://doi.org/10.1007/978-1-4419-9096-9
18. Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems*, 29. https://papers.nips.cc/paper/2016/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html
19. Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. *Proceedings of ICML*, 325–333. http://procee dings.mlr.press/v28/zemel13.html
20. Speicher, T., Heidari, H., Grunwald, T., & Krause, A. (2018). A unified approach to quantifying algorithmic unfairness. *Proceedings of the 35th International Conference on Machine Learning*. http://procee dings.mlr.press/v80 /speicher18a/speicher18a.pdf
21. Jiang, R., Pacchiano, A., Stepleton, T., Jiang, H., & Ma, T. (2020). Wasserstein fairness. *Proceedings of ICML*, 4160–4170. http://proceedings.mlr.press/v119/jiang20c/jiang20c.pdf
22. Bogen, M., & Rieke, A. (2020). *Help wanted: An examination of hiring algorithms, equity, and bias*. Upturn. https://www.upturn.org/reports/2020/hiring-algorithms/
23. Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW), 1–27. https://d oi.org/10.114 5/3392878
24. Tambe, P., Cappelli, P., & Yakubovich, V. (2019). Artificial intelligence in human resources management: Challenges and a path forward. *California Management Review, 61*(4), 15–42. https://doi.or g/10.1177/0 008125619 867910
25. Ghalwash, M., Raghavan, M., & Singh, M. (2021). Stability and fairness in learning-based workforce analytics. *Proceedings of AAAI*, 15268–15276. https://ojs.aaai.org/index.php/AAAI/article/view/17811
26. Duarte, J., Hu, Y., & Peng, L. (2020). Fairness in job classification models: An empirical study.
27. *Journal of Management Analytics, 7*(4), 593–610. https://doi.org/10.1080/23270012.2020.1806633
28. Van den Broek, E., Sergeeva, M., & Huysman, M. (2021). Hiring algorithms: Fairness, accountability, and transparency in organizational practice. *Journal of Business Ethics, 171*(4), 735–752. https://do i.org/10.1007/s10551-020-04463-y
29. Liem, C. (2021). Considerations for responsible algorithmic decision-making in HR. *Nature Human Behaviour, 5*(12), 1504–1513. https://doi.org/10.1038/s41562-021-01253-y
30. Willoughby, C., & Singh, L. (2022). Predictive attrition modeling: A systematic review. *Human Resource Development International, 25*(2), 213–235. https://doi.org/10.1080/13 678868.202 1.2017390
31. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpretable machine learning. *Advances in Neural Information Processing Systems*, 4765–4774. https://doi.org/10.48550/ arXiv.1705.078740
32. Molnar, C. (2022). *Interpretable machine learning*. Springer. https://doi.org/10.1007/978-3-030-65948-3