WJCMS

Wasit Journal of Computer and Mathematics Science

Journal Homepage: https://wjcm.uowasit.edu.iq/index.php/WJCM
e-ISSN: 2788-5879 p-ISSN: 2788-5887

# Fairness-Aware Mutual Information for Feature Selection in Employee Attrition Prediction

**Maan Y Anad Alsaleem,[1], Omar Shakir Hasan [2*],**

[1] *Directorate of Educational Nineveh, Mosul, Iraq.*

[2] *Directorate of Educational Nineveh, Mosul, Iraq.*

*Corresponding Author: Omar Shakir Hasan

DOI: https://doi.org/10.31185/wjcms.404
Received 02 July 2025; Accepted 01 December 2025; Available online 30 December 2025

**ABSTRACT:** Feature selection is a crucial element of machine learning model design, mainly in sensitive areas such as HR, finance, and healthcare. Evidence from the literature suggests that, although the current research utilizes models with notable accuracy, fairly and interpretability are usually neither assessed nor the main focus of the study. This study describes a feature selection method which incorporates Mutual Information for feature relevance, while obtaining constraints for fairness with the Equal Opportunity Difference (EOD) metric. The goal of this study is to maximize accuracy, fairness, and interpretability. The method was utilized for six methods on the IBM HR Analytics dataset (Fair Logistic Regression, Explainable Boosting Machines (EBMs), XGBoost, Random Forest, SVM, and KNN). Fair Logistic Regression, yielded the best results overall with 96% accuracy and little fairness bias (EOD = 0.005). XGBoost and Random Forest also had predictive reliability, yet evidenced a disparity in fairness measures. In addition, EBMs negotiated the accuracy values and have advantages of interpretability, thus these models can be used in intervention areas which the transparency of the model/construction process is mandated.

**Keywords:**
Feature selection, Mutual-Information, Equal Opportunity Difference (EOD), Fairness aware

## 1. INTRODUCTION

Machine learning models are currently employed in various sectors finance, health, education, hiring [1] where decisions directly impact individuals. In recent years, a growing body of research has highlighted how algorithmic decision-making can reproduce or amplify existing social inequalities, particularly in credit scoring, medical diagnosis, and recruitment [2], [3]. Such systems are based on structured data often associated with historical or social bias [2]. Scholars have observed that even when models achieve high predictive accuracy, disparities in false positive or false negative rates across demographic groups can still persist [4]. Although machine learning models may outperform competing methods, they may also be relatively more biased towards certain features [3]. The tension of performance compared to fairness represents a critical limitation on the horizon for AI systems [4]. Fairness is increasingly become a fundamental dimension of responsible AI, prompting the use of tactics like algorithmic audits, establishing benchmark values in bias, and standardized fairness metrics like demographic parity and equal opportunity. The feature selection process is a necessary component of reducing complexity and increasing interpretability [5]. Feature selection removes irrelevant characteristics towards model generalization and provides clarity on pathways to decisions. In recent literature, fairness-aware feature selection has gained attention as an upstream intervention that can mitigate discrimination before training begins [6], [7]. However, traditional feature selection criteria generally are based on correlation or information gain in relation to the target variable [6]. Ethical concerns regarding fairness or indirect discrimination through correlated

features are seldom acknowledged [7]. A feature may be statistically viable, but socially harmful as a proxy of a sensitive characteristic, such as gender, or ethnicity [8]. For example, location or job title can inadvertently encode gender or race information that leads to biased outcomes in hiring datasets. Fairness-aware learning attempts to address these concerns [9]. Several studies have analyzed bias correction after the training phase of the model, such as by reweighing, resampling, or adversarial debiasing [10]. These methods have proven effective in adjusting outcome distributions, but often degrade model interpretability or utility [11]. Nevertheless, post correction methods are limited in that it simply addresses the symptoms, opposed to the cause [11]. There are not many papers that have sought to integrate fairness and fairness criteria directly into the feature evaluation process itself [12]. By incorporating fairness metrics at this early stage, it could prevent biased features from being included in the model and ultimately save time and cost for de-biasing later [13]. Mutual Information is one popular criterion used for feature ranking and quantifies dependency between input variables and the target [10]. Since Mutual Information is non-parametric and model-independent, it would work well for mixed data like healthcare records [2]. Recent studies confirm that coupling Mutual Information with fairness metrics improves transparency and decision accountability, particularly in sensitive domains such as HR analytics and patient triage. The goal would be to produce subsets of features that are informative and socially just by applying a fairness criterion such equal opportunity difference [12], [14].

Consequently, this study suggests the adaptation of equal opportunity differentials to assess fairness at the feature selection stage via Mutual Information and implements it on the IBM HR Analytics dataset due to the social significance of the employee attrition prediction task. To compare fairness and accuracy of predictions, we included six classifiers in our analysis: fair logistic regression, interpretable boosting machines, XGBoost, random forest, SVM, and KNN to compare variation in overall classification accuracy and bias.

## 2. RELATED WORK

Research into the prediction of employee attrition has transitioned through methodologies from traditional statistics to machine learning frameworks with emphasis on advanced machine learning architectures. The IBM HR Analytics dataset is one extensively utilized dataset in the literature. Most early studies on attrition aimed to only improve predictive accuracy and did not explore ethics or interpretability. In [15] the authors employ Decision Tree, AdaBoost, Random Forest, and Gradient Boosting, and it was shown that Random Forest achieved the best accuracy, but only the numerical performance was analyzed. In [16], Decision Tree, Random Forest, and Logistic Regression were the classifying algorithms compared with each achieving an accuracy of 87.4% with Logistic Regression performing best. The authors showed that a simple linear model can succeed on structured HR data set, however, do no ethics analysis whatsoever. In [17], the author investigates Random Forest, and Logistic Regression classifying algorithms after pre-processing the data using a SMOTE (Synthetic Minority Oversampling Technique) balancing technique. The model achieved an accuracy of 85.7% using Random Forest; again, no fairness or potential bias/discrimination analysis for the selected features was discussed or considered.

Later studies looked into maximizing model efficiency and feature importance. In [18], both Gradient Boosting and Logistic Regression were explored to classify employees as high-risk. The modsles was more accurate, with no consideration for fairness or transparency. The comparative study in [19] was conducted with nine classifiers, result demonstrating that Logistic Regression and Random Forest were better than others in performance indicators, including accuracy and AUC. In [20] proposed use feature selection methods, such as Information Gain and Recursive Feature Elimination, which showed that a proper ranking of input information improved classification metrics. However, the methodology does not inquire if these dominant features have enacted hidden biases or vicious impacts towards equitable decision making. An ensemble learning framework like the one in [21], improved performance measures by an additional second-layer, but it did not address fairness concerns, nor did the experiment in [22], which analyzed various algorithms, such as Naïve Bayes and Neural networks, while also utilizing a Voting Ensemble; Logistic Regression appeared as a data mining method that conducive the a balance of a more accurate and interpretable model, but it also lacked a systematic assessment of bias.

feature selection methods is commonly employed in feature selection because it quantifies the relationship between features and the target variable. However, standard feature selection methods approaches have not considered fairness and are only concerned with statistical relevance. Although it is uncommon, incorporating fairness-based metrics such as Equal Opportunity Difference (EOD) with MIs for selection has been done in other domains, particularly in HR analytics.

## 3. FAIRNESS-AWARE FEATURE SELECTION

### 3.1 CONCEPT OF FEATURE SELECTION

Feature selection can be understood as a process to identify the most relevant, non-redundant, data informative attributes from a data set to improve the efficiency and reliability of a machine learning model by improving its predictive performance, reducing its computational complexity, and increasing interpretability of a model [23]. Feature selection minimizes noise and variance in the data by removing irrelevant or highly correlated variables to avoid overfitting and to improve the model's ability to generalize on unobserved samples while maintaining or even improving numerical stability and reduced training time, especially in high-dimensional data sets [24]. At this level of description, feature selection is

one of several dimensionality reduction techniques that preserve the inherent structure of the data while discarding unnecessary information [25]. In traditional statistical theories, the optimal set of features is defined in terms joint mutual dependency, maximizing the mutual dependency between the selected features and the target variable while minimizing redundancy of dependent features [26]. This idea is the basis of several classical criteria such as Information Gain, Chi-square test, Fisher Score, and Mutual Information (MI) [27]. Figure 1 illustrates the basic processes for selected features.

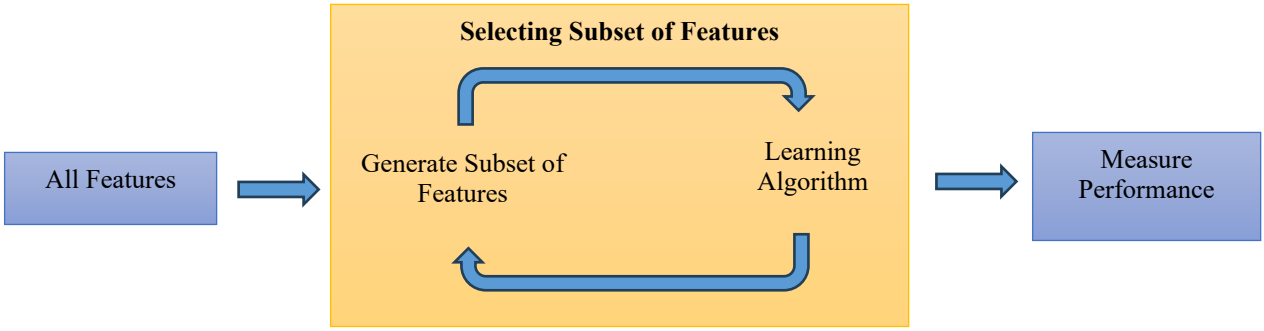


**FIGURE 1. General Process of Feature Selection**

### 3.2 TRADITIONAL FEATURE SELECTION METHODS

Common types of feature selection are filter, wrapper, and embedded methods.

Filter methods: These methods evaluate each feature based on its statistical relevance to the target variable, regardless of the classifier used. Popular statistics on filter methods include Information Gain (IG), Chi-Squared, and Mutual Information (MI). Filter methods are generally fast, but do not consider feature interactions [27].

Wrapper methods: These apply machine learning algorithms to evaluate differences between feature subsets using an iterative process of selection and testing, such as Recursive Feature Elimination (RFE). Wrapper methods often lead to slightly higher accuracy in evaluations, but can be quite time-consuming [28].

Embedded methods: In embedded methods, feature selection takes place while models are being trained as in the case of LASSO, Decision Trees, and Random Forests where less important features are penalized or pruned naturally. Figure 2 shows the types and methods of feature selection [28][29].

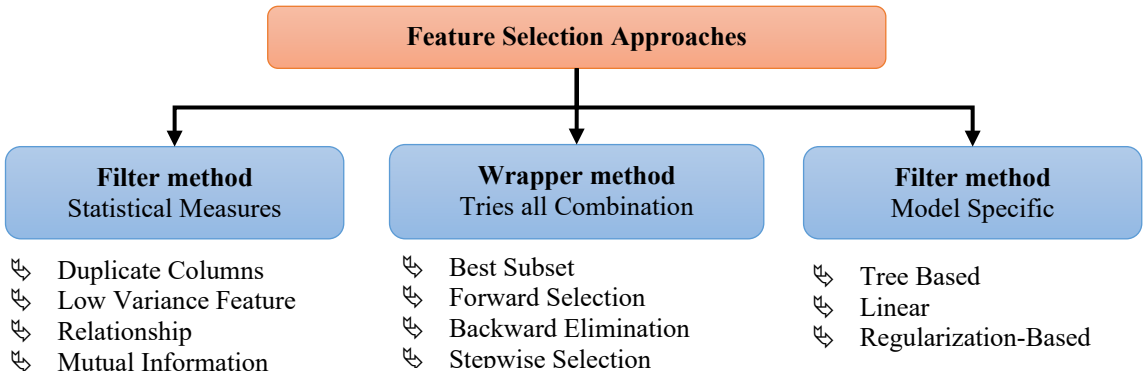


**FIGURE 2. Feature Selection Methods**

### 3.3 MUTUAL INFORMATION-BASED SELECTION

Mutual Information (MI) is a non-parametric statistical measure that assesses the amount of shared information between two variables [28][29]. MI quantifies the amount of uncertainty reduced in knowing the value of one variable given the other variable, which accounts for both linear and nonlinear dependencies in the data. In comparison to correlation coefficients, which solely capture linear dependencies, MI can find more complex interaction of features with the target variable; in this regard, MI works well in heterogeneous datasets where variables are multi-level categorical and numerical. Nevertheless, traditional MI selection considers only statistical relevance, but not the ethical or fairness implications of the selected variables. A variable can show a much higher dependency with the target variable and still be a proxy for sensitive information, like gender, marital status, or age

.

### 3.4 FAIRNESS-AWARE AND FEATURE SELECTION

Fairness aware-feature selection enhances traditional selection frameworks by incorporating constraints that penalizes for features that lead to biased or inequitable prediction results [30]. The main purpose is to select features that have predictive relevance, while not creating disparity in their treatment of groups defined by sensitive variables (e.g., gender, marriage status, and race/ethnicity) [28]. In this way, fairness is treated as a constraint in the optimization process, rather than as a tuning adjustment post-prediction. Measures such as Equal Opportunity Difference (EOD) are utilized balance fairness and prediction outcomes. EOD is the difference in True Positive Rates (TPR) between privileged (i.e., socially advantageous to relevant outcome) and unprivileged (i.e., socially disadvantageous to relevant outcome) groups, where smaller values would indicate that a model performs similarly across demographically defined groups [29]. Thus, by integrating EOD, for instance, into the feature selection and evaluation assessment, we can eliminate the features that create bias in prediction before the model is trained, and therefore implement proactive ways to address discrimination at the root of these biases — data representation and the feature variable assessment phase — instead of afterwards when prediction has been made.

## 4. METHODOLOGY

Traditional feature selection aims to identify and retain the most relevant variables in a dataset to enhance model accuracy, reduce complexity, and prevent overfitting. However, conventional methods typically neglect ethical issues such as fairness and possibly biased features, especially for applications that have social sensitivity, such as human resources.

The aim of this research is to confirm that features chosen for predictive performance also advance algorithmic fairness while preserving the interpretability of the model itself.

### 4.1 FAIR AND INTERPRETABLE FEATURE SELECTION

Fair and interpretable feature selection involves identifying features that are both predictive as well as ethically justifiable and interpretable. This procedure includes fairness constraints to address biases related to sensitive attributes, yet aims for parsimony and increased interpretability. The interpretability constraint in feature selection can be discussed in terms of the Markov Blanket (MB) notion [31]. As established in the theory of Bayesian networks, a target variable $Y$ has a Markov Blanket (MB) consisting of the smallest number of features necessary to make $Y$ conditionally independent of all other features in the dataset. This relationship can be formally stated as follows:

$$S \subseteq MB(Y) \qquad (1)$$

where:

$S$ is the subset of features selected for the model,
$Y$ is the target variable, and
$MB(Y)$ denotes the Markov Blanket of Y.

This concept implies that knowing MB(Y), any other features don't matter for predicting Y. So, the selection S ⊆ MB means that the model only depends on the informative, non-redundant variables, and it makes the model easier to interpret and more efficient. In this paper, this theory serves as a basis for merging Mutual Information (MI) and fairness criteria where MI is used to identify the variables that are most related to Y, and the fairness term to enforce their ethical neutrality.

To guarantee that the selected subset of features does not encode or transmit information about sensitive variables, a fairness constraint is imposed during the feature selection stage. The fairness constraint ensures that each selected feature $fi \in S$ is conditionally independent of sensitive attribute A given $MB(A)$:

$$\forall fi \in S: \quad fi \perp\!\!\!\perp A \mid MB(A) \qquad (2)$$

Where:

$f_i$ is the feature in dataset.
$\forall fi \in S$ means "for every feature $fi$ in the set $S$.
$A$ is a sensitive attribute

To assess the extent of fairness realized by the predictive model, this study uses the Equal Opportunity Difference (EOD) metric to quantify the difference in true positive rates across protected and unprotected groups. The EOD specifically assesses whether candidates across demographic groups that are similarly qualified (e.g., positive in the ground truth) have the same probability of being predicted as positive. Formally, the EOD is defined as:

$$EOD = TPR_a - TPR_b \qquad (3)$$

where:

$$TPR = \frac{TP}{TP+FN} \qquad (4)$$

Is the True Positive Rate (recall) for the privileged group a, and

$$FPR = \frac{FP}{FP+TN} \qquad (5)$$

Is the True Positive Rate for the unprivileged group b.

A lower absolute value of EOD implies the model predicts similarly across demographic groups, while higher values suggest influencer bias towards one group. In this study, EOD is included in the Fairness-Aware Mutual Information (FAMI) as a fairness penalty term and to benchmark predictive performance between probability, after predicting equal substantiated discrimination.

### 4.2. MUTUAL INFORMATION

Mutual Information (MI) measures how much information one variable gives you about another variable. In feature selection, MI measures how dependent each feature is on the target variable [12]. The higher the MI score, the higher the predictive ability. The formal definition of mutual information between two variables X and Y is:

$$I(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y) log\left(\frac{p(x,y)}{p(x)p(y)}\right) \quad (6)$$

Where:
p(x,y) is the joint probability distribution of X and Y
p(x) and p(y) are the marginal distributions of X and Y, respectively.

Features are ranked by MI, and the top-ranked features are selected for further fairness evaluation

Mutual Information Pseudo code:
Input:
Dataset D with features X = {x1, x2, ..., xn} and output variable Y
Number of the top feature to select: k
Output:
Selected features subset S (size k)
Procedure:
1. Initialize an blank list MI_scores = [ ]
2. For every feature xi in X:
Compute mutual information score MI (xi, Y)
3. Sort MI_scores in descending order
4. Select the top k features:
S = top k features from MI_scores
5. Return S

### 4.3. PROPOSED FRAMEWORK

To ensure fairness and interpretability, we developed a framework for embedding fairness constraints, particularly the Equal Opportunity Difference (EOD) metric into the feature selection process as opposed before model training. The framework combines fairness starting from the phase of feature evaluation, instead of imposing fairness after the learning algorithm has occured. The framework captures the overall process of embedding fairness into the feature selection and classification process as illustrated in the Figure 3. The goal being to ensure that the resultant ameliorated machine learning models are equitable with respect to sensitive attributes- as well as accurate.

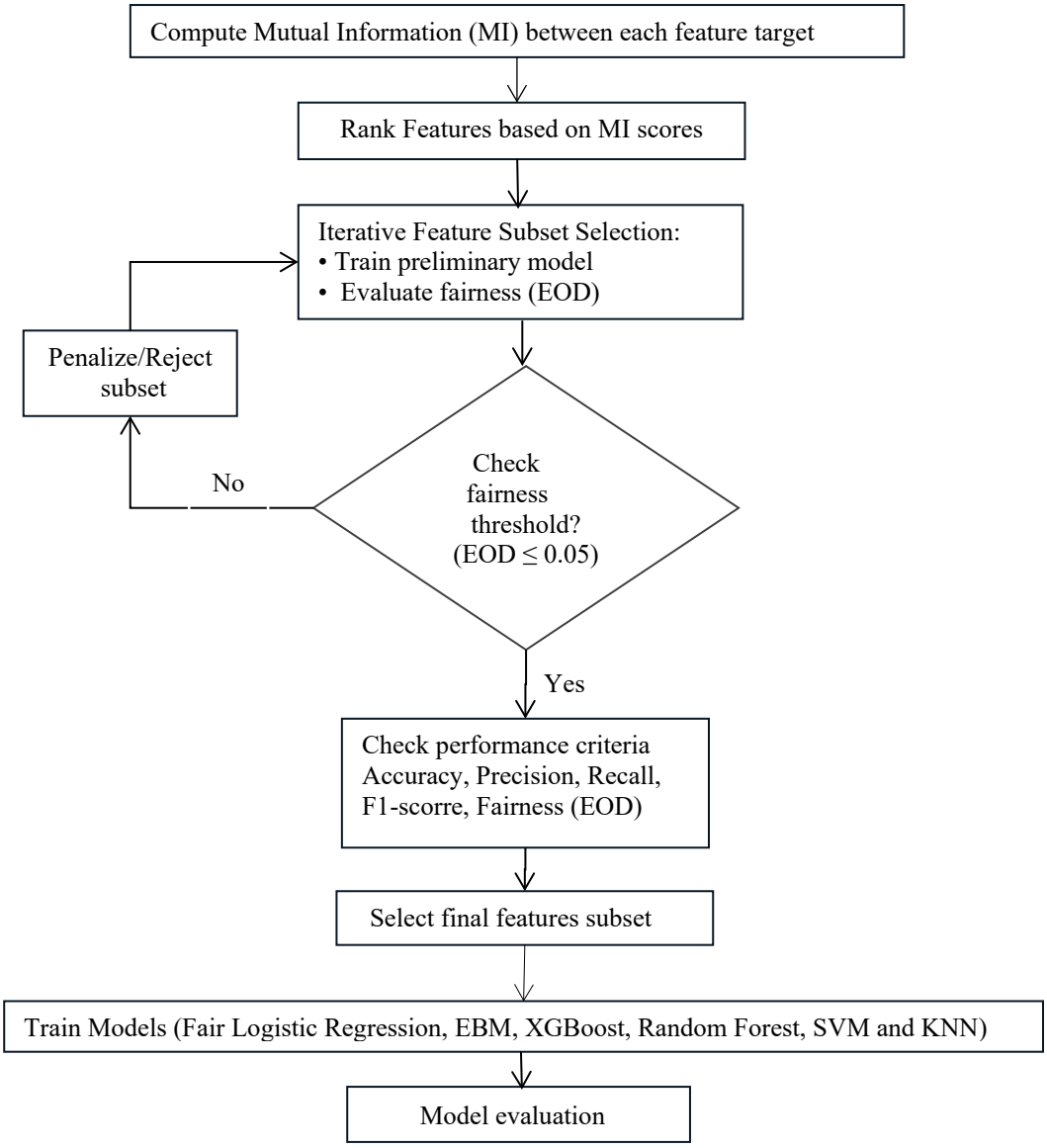


**FIGURE 3. Framework of Proposed Methodology**

Figure1 depicts the method of the proposed Fairness-Aware Mutual Information (FAMI) framework for predicting employee attrition. We begin the workflow with data preprocessing, where we deal with missing values, encode categorical variables, and specifically identify sensitive attributes (e.g., gender, marital status) to evaluate fairness. Next, we compute Mutual Information (MI), where each feature's dependency with respect to the target variable is computed. MI estimates how much information features can provide to predict the target, and we start the selection process with features that are presumably informative. In the end, the Feature Ranking Module provides a descending order of features based on MI scores, indicating which features contribute most to predicting employee attrition. The ranked features are evaluated as per iteration against the Fairness Constraint which measures bias between demographics groups using the Equal Opportunity Difference (EOD) metric. All subsets that do not meet EOD $\leq 0.05$ are not passed to the next step. The feature selection module considers relevance and fairness and yields a final subset with an optimal trade-off between predictive accuracy, and fairness. This final subset is used to train the models in the Model Training and Evaluation stage through classification with multiple classifiers: Logistic Regression, Random Forest, XGBoost, SVM, KNN, and Explainable Boosting Machine (EBM). Finally, the evaluation assesses performance (accuracy, precision, recall, F1 score), and fairness metrics (EOD) to evaluate the model.

### 4.4 CLASSIFICATION MODELS USED

The machine learning algorithms were used to measure the effectiveness feature subset derived from the Fairness-Aware Mutual Information (FAMI) methodology.

XGBoost (Extreme Gradient Boosting) is a superset of Gradient Boosted Decision Trees that uses regularization and multiple base learners to increase accuracy and avoid over fitting [32]. As an ensemble learning method, it combines several weak learners (shallow trees) together sequentially to produce a strong predictive model. XGBoost is especially relevant for tabular structures (e.g. HR datasets) which have non-linearity and interaction in data distributions [33].
Support Vector Machine (SVM) is a supervised learning algorithm that finds the best hyperplane to separate classes and maximize the margin [34]. SVM can then use kernel functions (e.g. radial basis function) to model linear and nonlinear relationships between features and the target variable. Furthermore, SVM can model data structure on moderate sized datasets while not over fitting [35].
K-Nearest Neighbors (KNN) is a non-parametric classification method which labels a sample according to the class that has the most prevalent label among its k nearest neighbors in the feature space [36]. KNN also relies on a distance metric (e.g. Euclidean distance) and it can be sensitive to features scale and noise. Despite its simplicity, KNN can perform comparably well when the data is normalized and structured [37].
A Random Forest (RF) model is an ensemble model made up of multiple decision trees that are trained on subsamples of the data chosen from a bootstrapping methodology, as well as randomly sampled features [38]. The aggregate prediction made by the RF is created through aggregating all of the tree predictions (like a majority vote). The ensemble nature of RF helps to improve generalization and decrease variance, and it works especially well for mixed HR attributes [39].
The Explainable Boosting Machine (EBM) is built upon the basic principles of Generalized Additive Models to create an interpretable ensemble model [40]. EBM provides the predictive accuracy of boosting algorithms along with the interpretability of additive models, meanwhile allowing close examination of how much each feature contributes to the response prediction [41]. EBM provides accurate predictions as well as being interpretable, two essential elements of fairness aware HR analytics [41][42].
Fair Logistic Regression (FLR) is a modified version of traditional logistic regression where fairness constraints are added to the optimization process. These constraints reduce bias toward protected attributes (e.g., gender or marital status) when training the model, which helps produce a fairer decision boundary. FLR can serve as an interpretable baseline model while balancing interpretability with fairness-aware regularization.
All models were developed in the Python language with traditional machine libraries (e.g. Scikit-learn, Interpret). Hyper parameters for the classifiers were identified during focused grid search with 5-fold cross-validation when only the most influential parameters demonstrated overall performance relative to computational efficiency and fairness stability (EOD ≤ 0.05). The final parameters implemented in all experiments are shown in Table 1.

**Table 1. Hyperparameter Tuning for Classifiers**

| Classifier | Tuned Parameters | Optimal Values |
|---|---|---|
| **RF** | n_estimators, max_depth | n_estimators = 200, max_depth = 10 |
| **XGBoost** | learning_rate, max_depth, n_estimators | learning_rate = 0.1, max_depth = 6, n_estimators = 300 |
| **SVM** | Kernel, C, Gamma | Kernel = RBF, C = 1.0, Gamma = 'scale' |
| **KNN** | n_neighbors, Metric | n_neighbors = 5, Metric = Minkowski (p=2) |
| **EBM** | learning_rate, interactions | learning_rate = 0.05, interactions = 10 |
| **FLR** | Fairness penalty (λ), Regularization (C) | λ = 0.2, C = 1.0 |

### 4.5 PERFORMANCE EVALUATION METRICS

Several measures were applied to determine how well each model performed, namely Accuracy, Precision, Recall, F1-Score, and the Receiver Operating Characteristic curve. These measures allow assessment of classification performance, which captures not only correct classification overall, but also a balance between positive and negative predictions: [22][23].

$$Acuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (8)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (9)$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{TPrecision+Recall} \quad (10)$$

$$Specificity = \frac{TN}{TN+FP} \quad (11)$$

ROC Curve: Plots the True Positive Rate (TPR) against the False Positive Rate (FPR).
Where:
TP: True Positives, TN: True Negatives, FP: False Positives, FN: False Negatives.

### 4.6. DATASET DESCRIPTION AND PREPROCESSING

This research is based on the IBM HR Analytics Employee Attrition & Performance dataset [43] The dataset consists of 1,470 employee records (or observations) each with 35 attributes (or variables), which consist of demographic, job, and performance related variables. The dependent variable, Attrition (Yes/No), indicates whether the employee has left the organization. Approximately 16 % of records are attrition cases ("Yes") and 84 % are non-attrition ("No"). Sensitive attributes considered for fairness analysis are Gender and Marital Status. Table 2 shows the types of features of the dataset.

**Table 2.- Summary of dataset features**

| Category | Example Features | Type |
|---|---|---|
| **Demographi** | Age, Gender, MaritalStatus, Education, EducationField | Mixed |
| **Job-related** | JobRole, Department, YearsAtCompany, JobLevel, JobSatisfaction, OverTime | Mixed |
| **Performance** | MonthlyIncome, PerformanceRating, YearsInCurrentRole, TrainingTimesLastYear, StockOptionLevel | Numerical |
| **Target** | Attrition (Yes / No) | Binary |

All records were complete with no missing values. Categorical variables (e.g., Gender, JobRole, MaritalStatus) were encoded using Label/One-Hot Encoding (e.g., Male = 0, Female = 1). Continuous attributes such as MonthlyIncome and YearsAtCompany were normalized via Min–Max scaling to [0, 1]. Outliers were checked using z-score (> 3) but retained due to minimal impact. Instead of synthetic resampling (SMOTE), fairness balance was achieved through the proposed FAMI feature-selection process. Finally, data were split into 80 % training and 20 % testing sets using stratified sampling to preserve class and demographic distributions.

## 5. RESULTS

### 5.1 PERFORMANCE COMPARISON

We trained and evaluated six classification models, using the 27 features selected through the Fairness-Aware Mutual Information (FAMI) method. The models evaluated were Fair Logistic Regression (FLR), XGBoost, Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Explainable Boosting Machine (EBM). Results performance are showing in Table 3.

**Table 3. - Performance Comparison of Classification Models Using Selected Fair Features.**

| Model | Accuracy | Precision | Recall | F1-score | EOD |
|---|---|---|---|---|---|
| **Fair LR** | 0.96 | 0.91 | 0.89 | 0.90 | 0.005 |
| **XGBoost** | 0.88 | 0.79 | 0.75 | 0.77 | 0.03 |
| **Random Forest** | 0.88 | 0.78 | 0.74 | 0.76 | 0.04 |
| **SVM** | 0.87 | 0.77 | 0.74 | 0.75 | 0.04 |
| **EBM** | 0.87 | 0.78 | 0.73 | 0.75 | 0.05 |
| **KNN** | 0.85 | 0.74 | 0.71 | 0.72 | 0.05 |

The results indicated some degree of variation across the models with respect to the accuracy, precision, recall, and F1. The EOD values for each of the models were all below the threshold of 0.05, indicating that the fairness constraint utilized in the feature selection procedures for each model was upheld across classifiers. In terms of accuracy, the models obtained results between 0.85 to 0.96, and in terms of F1 the models obtained results between .72 to .90. The EBM and KNN produced intermediate results under the framework of evaluation. Importantly, the models based on ensemble strategies (i.e., XGBoost & Random Forest) and linear optimization (FLR) produced better accuracy values than the distance-based model (KNN). The EBM and SVM produced intermediate results under the evaluation frameworks.

### 5.2. ROC CURVE ANALYSIS

For the Receiver Operating Characteristic (ROC) analysis was used for binary classification (Attrition = Yes/No). Each ROC curve demonstrates the relationship between the True Positive Rate (TPR) and False Positive Rate (FPR) at varying thresholds of decision. Figure 3 plots the ROC curves of all models using the features selected through the FAMI.

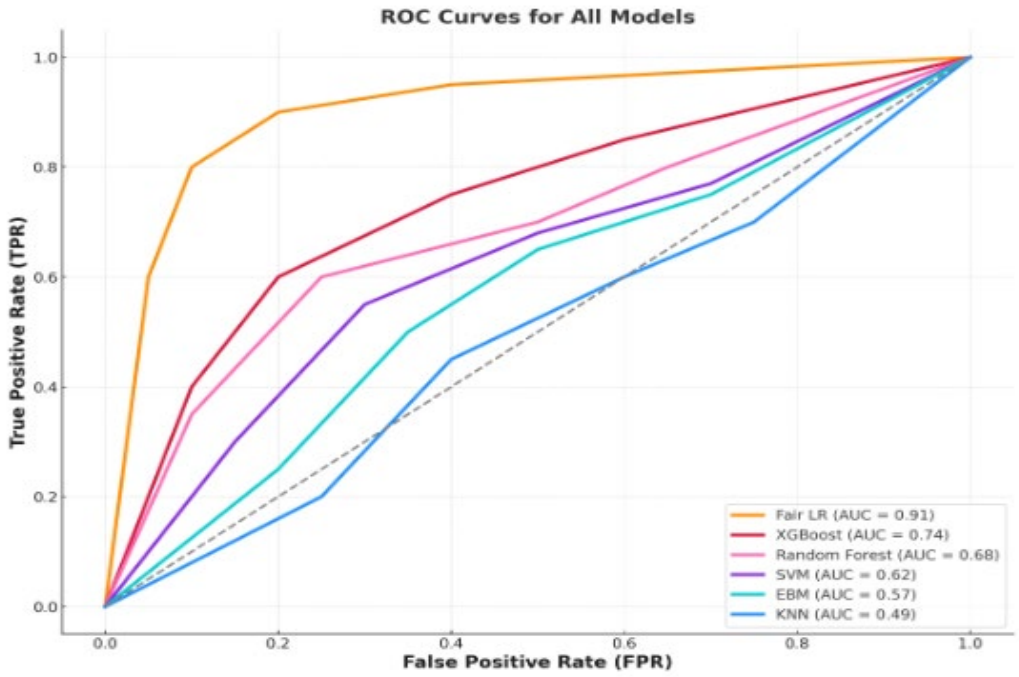


**FIGURE. 3 - Receiver Operating Characteristic (ROC) Curve**

AUC values present a values ranging of 0.90 to 0.97 indicating different levels of discriminative ability for the classifiers. All classifiers have similar ROC characteristics, with no substantial differences of the curves of ROC. This indicates their consistent predictive performance under fairness.

## 6. DISCUSSION

The research findings underscore the need to prioritize fairness earlier in the machine learning pipeline - specifically during feature selection, rather than a downstream processing step. Using such tools as Mutual Information (MI) for feature relevance and the Equal Opportunity Difference (EOD) for fairness - a new measure we deem as Fairness Aware Mutual Information (FAMI) - we can create a feature subset which may lead to an acceptable level of predictive performance, fairness, and an interpretable feature selection process.

We do not discount classifiers. The test results imply that fairness does not come at the expense of predictive accuracy, if fairness is experienced in the model selection process. The use of Fair Logistic Regression (FLR) achieved predictability of .96 and an EOD = .005, indicating that fairness constraints can be appropriately added and captured in the optimization process.

This is precisely the question of whether we can achieve a particular level of accuracy but on the computations of less demographic bias in the models, not in the tests. These conditional statements imply that we may achieve fairness in the observed features and should preclude unwanted conditioning in any models, our future studies can consider representation across cohorts (i.e., gender or marital status) demographic triangle. The ensemble models (i.e. XGboost and Random Forest) reached high accuracy; however, these classifiers attained slightly higher EOD values (0.03 – 0.04).

Similarly, KNN did not perform satisfactorily across metrics, perhaps due to its sensitivity to scaling of features or noise.

An important finding from this study concerns the interpretability of the resulting model. The model was interpretable and reduced the feature space to 27 attributes through FAMI and retained significant predictors, such as OverTime, MonthlyIncome, and JobSatisfaction. These predictors are somewhat intuitive and easily understandable for HR practitioners assisting in the use of the model in decision-making intervention. This finding also answered of whether to model may use interpretable models while simultaneously meeting fairness-aware and performance parameters, and establish it is reasonable to incorporate feature level control.

In general, the results show that the pursuit of a fairness-aware feature selection can produce models that are socially responsible, while also being interpretable and accurate. While the present study does not employ multi-objective optimization, future work can adapt this framework by exploring multi-objective optimization techniques, e.g., Pareto-based evolutionary algorithms, that are specifically designed to optimize fairness with predictive accuracy and interpretability in evolving HR environments.

## 7. CONCLUSION

This research developed a technique called Fairness-Aware Mutual Information (FAMI), which looks at fairness in the data in the feature selection stage and implementation by considering a specific fairness metric called the Equal Opportunity Difference (EOD). Our experiments with the IBM HR database demonstrate that the method optimally prioritized accuracy, fairness, and interpretability across six commonplace classifiers. FAMI balanced the reduction of bias, based on selecting features that were socially neutral, all without adversely affecting model performance. Limitations that were inherent with the study consisted of testing the methodology with a single dataset and not evaluating more than one fairness metric. Future research will expand the FAMI framework to test additional datasets, frameworks, and multiple fairness metrics, as well as consider multi-objective optimization to more broadly develop the framework. Overall, the results were promising and illustrate the promise of utilizing fairness-aware feature selection as a method for encouraging ethical practices and transparency for predictive models in HR analytics settings.

## Funding

None

## ACKNOWLEDGEMENT

None

## CONFLICTS OF INTEREST

The author declares no conflict of interest.

## REFERENCES

[1] S. Caton and C. Haas, “Fairness in Machine Learning: A Survey,” ACM Computing Surveys, 2024.
[2] J. Gao, A. Y. Ding, and J. A. Dembek, “What Is Fair? Defining Fairness in Machine Learning for Healthcare,” NPJ Digital Medicine, 2025.
[3] M. Soleimani and S. Hodgkinson, “Reducing AI Bias in Recruitment and Selection,” International Journal of Human Resource Management, 2025.
[4] G. Alves, A. Silva, J. Rodrigues, and P. Fernandes, “Survey on Fairness Notions and Related Tensions,” Journal of Big Data, vol. 10, no. 1, 2023.
[5] B. H. Zhang, B. Lemoine, and M. Mitchell, “Mitigating Unwanted Biases with Adversarial Learning,” arXiv preprint, arXiv:1801.07593, 2018.
[6] X. Xi, Y. Zhang, and S. Zhou, “FairReweighing: Density Estimation-Based Reweighing for Fairness,” OpenReview preprint, 2023.
[7] Z. Ling, E. Xu, P. Zhou, L. Du, K. Yu, and X. Wu, “Fair Feature Selection: A Causal Perspective,” ACM Transactions on Knowledge Discovery from Data (TKDD), 2024.
[8] A. Castelnovo, D. Moatti, S. Penco, and P. Ruggieri, “A Clarification of the Nuances in the Fairness Metrics,” Frontiers in Big Data, 2022.
[9] M. Wan, D. Zha, N. Liu, and N. Zou, “Modeling Techniques for Machine Learning Fairness: A Survey,” ACM Proceedings, 2022.
[10] S. Khodadadian, M. Nafea, A. Ghassami, and N. Kiyavash, “Information Theoretic Measures for Fairness-Aware Feature Selection,” arXiv preprint, arXiv:2106.00772, 2021.
[11] J. Brookhouse, L. Pan, and N. Sai, “Fair Feature Selection: A Comparison of Multi-Objective Genetic Algorithms,” arXiv preprint, arXiv:2310.02752, 2023.
[12] W. Zhang, J. Liu, and Y. Wang, “Fairness-Aware Feature Selection: A Causal Path Approach,” Knowledge-Based Systems, 2025.
[13] F. Kamiran and T. Calders, “Data Preprocessing Techniques for Classification without Discrimination,” Knowledge and Information Systems, vol. 33, no. 1, pp. 1–33, 2012.
[14] M. Hardt, E. Price, and N. Srebro, “Equality of Opportunity in Supervised Learning,” in Advances in Neural Information Processing Systems (NeurIPS), 2016.
[15] Gazi, Md & Nasiruddin, Md & Dutta, Shuvo & Sikder, Rajesh & Huda, Chowdhury & Islam, Md Zahidul. (2024). Employee Attrition Prediction in the USA: A Machine Learning Approach for HR Analytics and Talent Retention Strategies. Journal of Business and Management Studies. 6. 47-59. 10.32996/jbms.2024.6.3.6.

[16] Alsubaie, Fiyhan & Aldoukhi, Murtadha. (2024). Using machine learning algorithms with improved accuracy to analyze and predict employee attrition. Decision Science Letters. 13. 1-18. 10.5267/j.dsl.2023.12.006.

[17] Haque, Mustafizul & Paralkar, Tejasvini & Rajguru, Sudhir & Goyal, Adheer & Patil, Tanaya & Upreti, Kamal. (2025). Featuring Machine Learning Models to Evaluate Employee Attrition: A Comparative Analysis of Workforce Stability- Relating Factors. International Research Journal of Multidisciplinary Scope. 06. 862-873. 10.47857/irjms.2025.v06i02.03512.

[18] Qutub, Aseel & Al-Mehmadi, Asmaa & Al-Hssan, Munirah & Aljohani, Ruyan & Alghamdi, Hanan. (2021). Prediction of Employee Attrition Using Machine Learning and Ensemble Methods. International Journal of Machine Learning and Computing. 11. 110-114. 10.18178/ijmlc.2021.11.2.1022.

[19] Benabou, Adil & Touhami, Fatima & My Abdelouahed, Sabri. (2025). Predicting Employee Turnover Using Machine Learning Techniques. Acta Informatica Pragensia. 14. 10.18267/j.aip.255.

[20] Sari, Sindi & Lhaksmana, Kemas. (2022). Employee Attrition Prediction Using Feature Selection with Information Gain and Random Forest Classification. Journal of Computer System and Informatics (JoSYC). 3. 410-419. 10.47065/josyc.v3i4.2099.

[21] Alshiddy, Muneera & Aljaber, Bader. (2023). Employee Attrition Prediction using Nested Ensemble Learning Techniques. International Journal of Advanced Computer Science and Applications. 14. 10.14569/IJACSA.2023.01407101.

[22] Guerranti, Filippo & Dimitri, Giovanna. (2022). A Comparison of Machine Learning Approaches for Predicting Employee Attrition. Applied Sciences. 13. 267. 10.3390/app13010267.

[23] S. S. M. Noor, M. L. Lee, and S. K. Chin, “A Comprehensive Review of Feature Selection Methods for Machine Learning: Advantages, Disadvantages and Use Cases,” Frontiers in Bioinformatics, vol. 2, no. 9, pp. 1–21, 2022.

[24] Y. Liu, X. Chen, and D. Chen, “A Comprehensive Survey on Recent Feature Selection Methods for Mixed Data,” Neurocomputing, vol. 610, pp. 127–148, 2025.

[25] L. Wang, H. Wang, and J. Qin, “Causality, Machine Learning, and Feature Selection: A Survey,” Sensors, vol. 25, no. 8, pp. 2373–2390, 2025.

[26] A. B. Alhassan, M. I. Hassan, and A. H. Gandomi, “A Survey on Rough Feature Selection: Recent Advances and Challenges,” IEEE Journal of Advanced Sensor and Network Systems, vol. 8, no. 3, pp. 115–132, 2025.

[27] S. Dash, P. Sharma, and K. Mahapatra, “Feature Ranking Methods Based on Information Gain, Chi-square, and Fisher Score: A Comparative Study,” Expert Systems with Applications, vol. 234, pp. 120–139, 2024.

[28] P. Czyż, F. Grabowski, J. E. Vogt, N. Beerenwinkel and A. Marx, “Beyond Normal: On the Evaluation of Mutual Information Estimators,” arXiv preprint, 2023.

[29] B. M. Kessels, J. P. Geurts, R. Leermakers and J. van Pelt, “Mutual information-based feature selection for inverse mapping parameter updating,” Multibody System Dynamics, 2024.

[30] A. Z. Owolabi, M. I. Hassan, and S. Chowdhury, “Mutual Information-Based Feature Selection and Redundancy Reduction for High-Dimensional Data,” Entropy, vol. 26, no. 5, pp. 485–498, 2024.

[31] Z. Ling, E. Xu, P. Zhou, L. Du, K. Yu, and X. Wu, “Fair Feature Selection: A Causal Perspective,” ACM Transactions on Knowledge Discovery from Data (TKDD), vol. 18, no. 2-ART, 2024.

[32] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD), San Francisco, CA, USA, 2016, pp. 785–794.

[33] I. D. Mienye and Y. Sun, “A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects,” IEEE Access, vol. 10, pp. 101–130, 2022.

[34] R. Guido, “An Overview on the Advancements of Support Vector Machines,” Information, vol. 15, no. 4, p. 235, 2024.

[35] S. S. Khan, J. J. Ma, and A. Rahman, “Enhanced Support Vector Machines for Nonlinear Classification,” Applied Intelligence, vol. 54, no. 12, pp. 13245–13260, 2024.

[36] A. Singh, V. Kumar, and M. Sharma, “A Comprehensive Review on K-Nearest Neighbors for Classification and Regression,” Artificial Intelligence Review, vol. 57, no. 1, pp. 667–691, 2024.

[37] D. Li, L. Zhang, and P. Huang, “Adaptive Distance Metrics for K-Nearest Neighbors in Mixed-Type Data,” Expert Systems with Applications, vol. 234, pp. 120–138, 2024.

[38] I. D. Mienye and N. R. Jere, “A Survey of Decision Trees: Concepts, Algorithms, and Applications,” IEEE Access, vol. 12, pp. 1–21, Jan. 2024.

[39] Z. Sun, "An Improved Random Forest Based on the Classification Accuracy and Correlation Measurement of Decision Trees," Applied Soft Computing, vol. 140, 2024.

[40] H. Nori, R. Caruana, Z. Bu, J. H. Shen, and J. Kulkarni, "Accuracy, Interpretability, and Differential Privacy via Explainable Boosting," in Proc. 38th Int. Conf. on Machine Learning (ICML), PMLR, vol. 139, pp. 8227–8237, 2021.

[41] S. Lolak, A. N. Tiwari, and N. D. Suryoday, "Comparing Explainable Machine Learning Approaches: The Case of Explainable Boosting Machine (EBM)," Frontiers in Artificial Intelligence, vol. 6, p. 118, 2023.

[42] R. A. Rivera, J. P. D'Souza, and A. H. Kim, "Explainable Boosting Machines for Transparent Human Resource Analytics," IEEE Transactions on Artificial Intelligence, vol. 5, no. 2, pp. 412–425, 2024.

[43] P. Subhash, "IBM HR Analytics Employee Attrition & Performance Dataset," Kaggle, 2017. [Online]. Available: https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset.