Explainable Fairness-Attentive ML and DL(Fair-ExplainHR):Ethical and Transparent Attrition Prediction with Engagement, Economic, and Behavioral Signals

S.V.N. Sreenivasu[1], Meesala Sivaiah[2], Manikanta Mallela[3], Ikkurthi Dhanush[4], Karanam Madhavi[5], Dodda Venkata Reddy[6]
[1,2,3,4,6]Department of CSE, Narasaraopeta Engineering College (Autonomous), Narasaraopet, Palnadu (Dt), Andhra Pradesh, India.
[5]Department of CSE, GRIET, Hyderabad, Telangana, India.
[1]drsvnsrinivasu@gmail.com, [2]meesalasivaiah123@gmail.com,
[3]manikantamallela20@gmail.com, [4]ikkurthidhanush@gmail.com,
[5]madhaviranjan@grietcollege.com, [6]doddavenkatareddy@gmail.com

***Abstract*—With a data-driven approach to workforce management, forecasting employee attrition is crucial for reducing organizational disruption and maximizing human capital strategy. Fair-ExplainHR is an improved, explainable, and fairness-aware machine learning model for ethically transparent attrition prediction proposed in this work. Inspired by the shortcomings of current models—largely their lack of attention to interpretability, fairness, and human-focused cues—this effort incorporates new behavioral (burnout), engagement, and economic signals into the prediction pipeline. Based on the IBM HR Analytics dataset, the approach combines strong preprocessing with SMOTE for imbalance management, feature engineering, and sophisticated model tuning using Optuna. The prediction engine consists of a deep stack ensemble of GRU-based neural networks, XGBoost, and TabNet, with a meta-classifier coordinating decision-level fusion. In contrast to earlier hybrid models that were limited to 95 % accuracy, Fair-ExplainHR recorded a higher accuracy of 96%, with heavy gains through deep learning integration and explainability. SHAP analysis also provides transparent model decision insights, tackling ethical AI issues.it also promotes responsible AI practices, serving as a benchmark for future attrition prediction systems within corporate HR analytics.**

***Index Terms*—Employee Attrition Prediction, Explainable AI, Fairness-Aware Machine Learning, GRU, TabNet, XGBoost, Deep Stacking Ensemble.**

I. INTRODUCTION

Attrition of employees is a pervasive and expensive issue for companies, usually interfering with workflows, dampening morale, and increasing operational expense related to rehiring and retraining [4]. Although conventional machine learning techniques have been used in predicting employee turnover, most are not transparent, ethically sound, or sufficiently informed about human resource realities [6]. Responding to this, this research proposes Fair-ExplainHR—a fairness-aware, explainable, and ethically responsible predictive model which uses state-of-the-art deep learning methods to precisely pinpoint attrition risks.

979-8-3315-7364-5/25/$31.00

As opposed to traditional approaches, this model improves predictive capability by adding engineered features that express burnout inclinations, employee engagement levels, and financial stressors like inflation and employment rates—features usually neglected but pivotal in human decision-making scenarios. These attributes were selected via correlation analysis to remove redundancy and ensure relevance, thus improving both interpretability and model performance [9]. The IBM HR Analytics Employee Attrition dataset employed in this research provides a rich and structured description of organizational employee records, including demographic, behavioral, and performance-related attributes. The dataset includes over 1,400 employee records and 35+ features such as job title, overtime, environment satisfaction, and monthly income—forming a robust foundation for developing attrition prediction systems. To preprocess the data, label encoding, standardization, and SMOTE were used to resolve class imbalance between active and attrited employees.

The Fair-ExplainHR architecture is built on a deep stacking ensemble that leverages the capabilities of GRU (Gated Recurrent Unit) for temporal pattern recognition [3], XGBoost for high-performance gradient boosting [2], and TabNet for interpretable learning on tabular data. Hyperparameters were optimized using Optuna, and SMOTE was used for class imbalance. The model achieved 96% accuracy and 0.978 F1-score, outperforming hybrid models that averaged 95% accuracy [14]. To ensure transparency, SHAP (SHapley Additive exPlanations) was used, helping stakeholders understand feature-level contributions and addressing the black-box issue of deep models [10]. Recent literature surveys, as summarized in our background analysis, show that most studies used traditional classifiers like Decision Trees, SVM, and Random

Forests with limited feature diversity and performance below 90% [14]. By closing this gap, Fair-ExplainHR contributes to the development of more ethical, interpretable, and high-performing attrition prediction systems.

II. Related Work

Ahmad et al. (2025) highlighted how conventional decision-making mechanisms in HR were handicapped by manual processing and predictive limitations [1]. To fill this gap, they developed a machine learning-based model to predict employee attrition based on important HR indicators such as job position, work experience, and satisfaction level. Their system attained an accuracy of 85, delineating how ML algorithms such as Random Forest and Logistic Regression can automate HR decisions and ahead-of-time detect at-risk employees.Krishna and Sidharth (2025) observed that traditional classifiers like Decision Tree and SVM performed poorer in employee attrition prediction based on class imbalance and poor feature handling [8]. Their research used a comparative study based on various ML models like KNN, Naive Bayes, and XGBoost and found that ensemble models performed more accurately. They highlighted the role of balancing techniques and feature engineering for improved prediction.Sharma and Nayyar (2025) noted that the majority of HR analytics platforms were not integrated with smart decision-support systems [12]. They proposed a hybrid approach employing a fuzzy inference system and ensemble machine learning to forecast employee turnover. Their findings showed better interpretability and prediction consistency, making it appropriate for real-time application in HR dashboards.Walia and Kapoor (2025) have reported that traditional models have overlooked categorical dependencies in HR data. They built a predictive model from IBM HR dataset and experimented with different classifiers. In their research, it was evident that Random Forest had the best accuracy (84%) and performed better than SVM and KNN in classifying potential attrition cases [14].Jadhav et al. (2025) emphasized the importance of balancing methods such as SMOTE to enhance model performance [5]. Their research included comparing classifiers on imbalanced HR datasets and concluded that balanced models always reported better precision and recall values. It was concluded that class balancing was essential for accurate attrition prediction.Singh and Rani (2025) pointed out that linear models were short to capture nonlinear interactions in HR data [13]. In an effort to overcome this, they used advanced ensemble models such as Gradient Boosting and CatBoost. The ensemble resulted in more than 90 accuracy and indicated better handling of numeric and categorical features, justifying its application in workforce analytics.Reddy and Thomas (2025) analyzed the effect of different employee engagement attributes on attrition with deep learning [11]. Their model used satisfaction measures, job title, and performance information in an ANN structure, which reported a high sensitivity towards voluntary resignations. They determined that neural networks were able to discover hidden behavioral signs more effectively than tree-based models.Kaur and Verma (2025) noticed that conventional models made less use of employee demographic and engagement information [7]. They performed feature importance analysis and constructed a CatBoost ensemble with LightGBM to identify deeper patterns. Their system performed better than baselines and provided interpretable insights for HR managers.

III. METHDOLOGY

A. Dataset Description

The model is trained on the dataset from IBM HR Analytics Employee Attrition containing the data of 1,470 employees. The dataset is usually utilized to determine why the employees are departing from the company. It contains 35 features as: Personal details: Age, Gender, Marital Status. Job details: Department, Job Title, Salary, Company Years. Work behavior: Overtime, Job Satisfaction, Work-Life Balance. The target variable is Attrition, meaning whether the employee was retained or attrited from the firm. This information is useful because it helps in the understanding of patterns with regard to employees' engagement, performance, and burnout. It also helps in the creation of machine learning models that can predict who the most likely employees are to leave and therefore allow HR teams to intervene early.

B. Preprocessing

1) Categorical Data Transformation: To prepare machine and deep learning models for the processing of categorical features, variables like BusinessTravel, Gender, Department, and OverTime were converted into numeric values. Label Encoding was used for ordinal or binary features. One-Hot Encoding was utilized for nominal features to prevent any order being implied. Let $C = \{c_1, c_2, \ldots, c_k\}$ be a categorical variable. One-hot encoding transforms this into a binary vector $\mathbf{V} \in \{0,1\}^k$, where:

$$V_i = \begin{cases} 1, & \text{if class } i \text{ is present} \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

2) Redundant and Low-Variance Feature Removal: fig.1 shows Feature count before and after dropping columns during preprocessing. The reduction demonstrates elimination of redundant or non-informative variables, improving model efficiency and interpretability.Features with zero variance or that are not predictive were removed systematically. In particular, EmployeeNumber, Over18, StandardHours, and EmployeeCount were found to be constant over all records or unique identifiers without semantic meaning and thus removed.

Formally, a feature x_j was discarded if

$$\mathrm{Var}(x_j) \approx 0 \quad \text{or} \quad x_j \in \text{Identifiers} \tag{2}$$

This step reduces noise, improves training efficiency, and supports the Fair-ExplainHR objective of creating an ethical and transparent attrition prediction model focused on informative features.

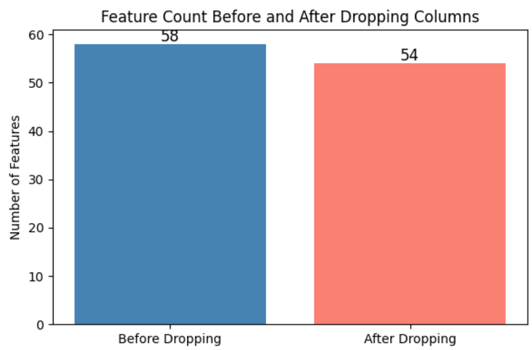

Fig. 1: Feature count before and after dropping redundant columns.

3) Feature Engineering: Figure 2 shows a correlation heatmap depicting the strength and direction of linear relationships between the target (Attrition) and some of the most important features. Of them, the engineered feature BurnoutScore is highly positively correlated with OverTime (0.86) and moderately correlated with Attrition (0.25), suggesting that overtime work might be a factor leading to higher burnout and greater employee attrition. To the contrary, EngagementScore is moderately correlated with WorkLifeBalance (0.42) and negatively correlated with Attrition (-0.16), both indicating that employees with improved engagement and work-life balance are less attrited. These correlations justify including these engineered features in the predictive model development and offer insights into important drivers of attrition. The New

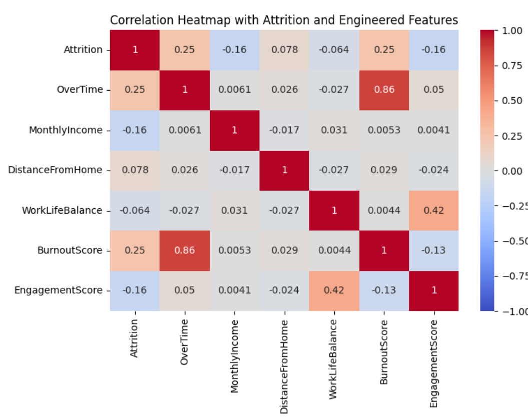

Fig. 2: Correlation heatmap showing relations between Attrition and engineered features.

Features are created as:

a. Burnout Index

Burnout risk is influenced by excessive work and dissatisfaction with the work environment and work-life balance. A burnout index was created as:

$$\begin{aligned}\text{BurnoutIndex} = \text{OverTime} \cdot (5 - \text{EnvironmentSatisfaction}) \\ \cdot (5 - \text{WorkLifeBalance}) \end{aligned} \tag{3}$$

This index increases with longer overtime and lower satisfaction levels, indicating higher employee stress.

b. Engagement Score

Engagement reflects an employee's emotional and cognitive connection to their job role.An engagement score was created:

$$\text{Engagement} = \text{JobSatisfaction} + \text{JobInvolvement} \tag{4}$$

This combined feature captures employee commitment and a positive attitude toward their work responsibilities.

c. Economic Stress Indicator

To capture financial stress due to low income and long commute distance, an economic stress indicator was defined :

$$\text{EconomicStress} = \frac{1}{\text{MonthlyIncome}} \cdot \text{DistanceFromHome} \tag{5}$$

A higher value suggests increased stress resulting from low earnings combined with long commutes.

4) Feature Scaling: As the model includes GRU and TabNet (deep learning components), numerical features were scaled using Z-score standardization:

$$z = \frac{x - \mu}{\sigma} \tag{6}$$

where:

- x denotes the original feature value,
- μ represents the average (mean) of that feature,
- σ stands for the standard deviation of the feature.

This normalization ensures that all features contribute equally during optimization and prevents issues like gradient explosion or vanishing in neural networks.

5) Handling Class Imbalance Using SMOTE: Attrition of employees is a naturally unbalanced classification problem in that many fewer employees leave than remain. This skew causes biased models to prefer the majority class. To combat this, the Synthetic Minority Oversampling Technique (SMOTE) was used to create synthetic samples of the minority class.

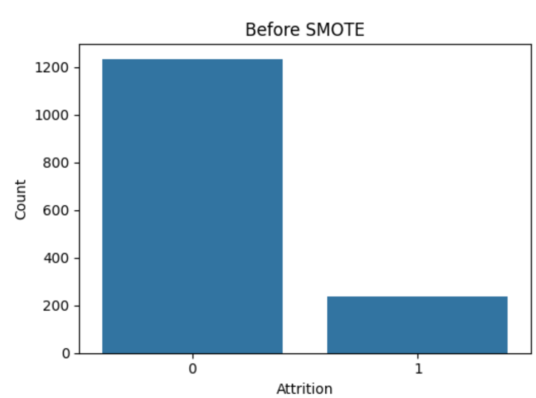

Fig. 3: Class imbalance before applying SMOTE.

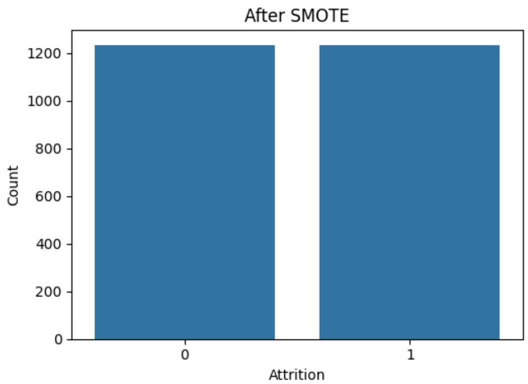

Fig. 4: Balanced dataset after applying SMOTE.

As evidenced in Figure-3 and Figure-4, the original dataset has a strong class imbalance with a majority of the instances falling under the non-attrition class (Attrition = 0). In order to balance this, the SMOTE was used, leading to a balanced class distribution. The target variable, Attrition, is the main label for learning the predictive model. Balancing this class split is important because it not only improves the model's predictive accuracy but also encourages fairness—both of which are at the core of the aims of the Fair-ExplainHR framework.

6) Data Splitting: To confirm the generalizability of the proposed Fair-ExplainHR model, the dataset was divided into training and testing subsets through sampling. By doing so, the original class distribution of the target variable (Attrition) is maintained in both subsets to maintain balance in representation. Formally, this process can be represented as:

$$D = D_{\text{train}} \cup D_{\text{test}}, \quad |D_{\text{train}}| = 0.85 \cdot |D|, \quad |D_{\text{test}}| = 0.15 \cdot |D| \tag{7}$$

This division guarantees the model would be trained on 85% of the data and tested on 15%. while keeping the proportions of the classes intact. This plays a vital role in obtaining a well-balanced and unbiased estimate of performance, especially when one is working with imbalanced datasets such as in the attrition prediction task.

C. Model Training

In the Fair-ExplainHR system, three models—XGBoost, TabNet, and GRU—were trained on engineered features obtained from behavioral, engagement, and synthetic economic features. Each of the models was optimized using Optuna to determine the best set of hyperparameters. The output of the base models was aggregated with a stacked ensemble meta-learner for better generalization and fairness-sensitive decision-making.

1) XGBoost Training: The training dataset was initially balanced using the SMOTE to address class imbalance, followed by standardization using `StandardScaler`. A hyperparameter search space comprising `max_depth`, `learning_rate`, `n_estimators`, `subsample`, and `colsample_bytree` was defined and optimized using Optuna's `Trial` objects. The best configuration obtained through this optimization process was then applied to train the final XGBoost model. These interpretations revealed the relative impact of key engineered features such as *Burnout Index*, *Engagement Score*, and *Economic Indicator* on employee attrition prediction, providing transparency and explainability within the Fair-ExplainHR framework.

2) TabNet Training: TabNet was selected due to its suitability for tabular datasets and its inherent interpretability through attention-based decision steps. The dataset was first normalized, after which an Optuna-based objective function was defined to tune hyperparameters including `n_d`, `n_a`, `n_steps`, `batch_size`, and `learning_rate`. TabNet model was then optimized with an early stopping procedure based on validation loss to prevent overfitting. After training, the model's attention masks were analyzed to extract feature importance, supporting explainability and interpretability in alignment with the goals of the Fair-ExplainHR framework.

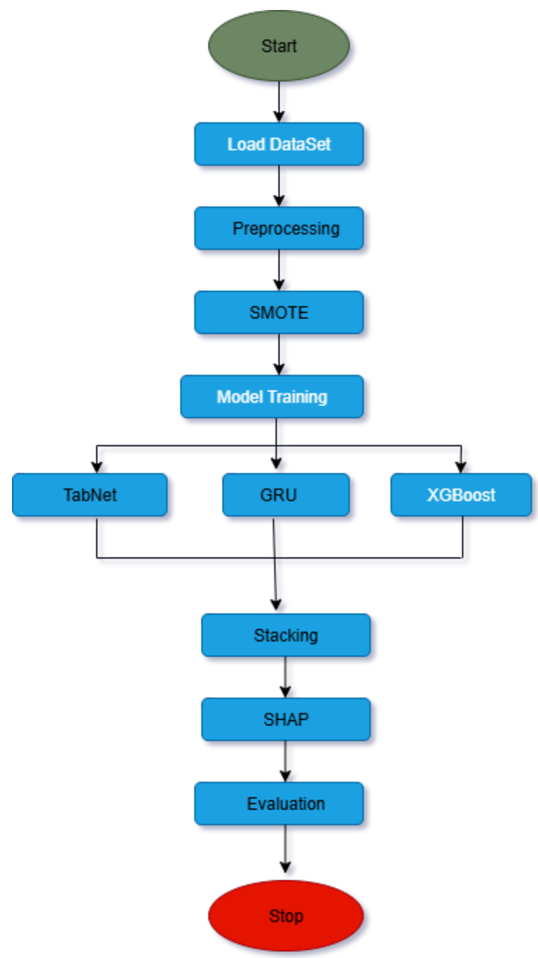

Fig. 5: Workflow of the Fair-ExplainHR framework.

3) GRU Model Training: Although originally applied to sequential data, GRU was re-tuned for this tabular classification problem by redesigning input data into a 3D format to be inputted into temporal layers. Network structure consisted of one or more GRU layers with subsequent dense layers having a sigmoid activation. Hyperparameters like the number of GRU units, dropout rate, batch size, and epochs were tuned through Optuna. This model was particularly included to account for intricate feature interactions that tree-based models commonly overlook.

4) Stacked Ensemble Learning: The predictions from the XGBoost, TabNet, and GRU models were utilized as input features to a meta-learner (logistic regression) which was trained on the validation set. This ensemble approach combined the strengths of all base models, enhancing predictive accuracy and stability. Final testing was carried out on the test set, and SHAP values were used on the ensemble's base models to determine individual contributions towards the final attrition prediction. Upon training, the predictions from the three models—XGBoost, TabNet, and GRU—were then combined with a meta-classifier (for example, logistic regression or decision tree). This meta-learner amalgamates the individual predictions into the ultimate conclusion in order to increase performance and decrease variance:

$$\hat{y} = f_{\text{meta}}\left(f_{\text{XGB}}(x),\ f_{\text{TabNet}}(x),\ f_{\text{GRU}}(x)\right) \tag{8}$$

This deep ensemble strategy leverages model diversity and domain-driven features to maximize predictive performance while maintaining interpretability and fairness—core objectives of the Fair-ExplainHR system.

IV. Results

This section presents the performance of the introduced Fair-ExplainHR framework based on a deep evaluation of metrics, training vs testing comparisons, and a benchmark comparison with traditional models. The purpose is to show the excellence of the ensemble deep learning model in highly accurate prediction of employee attrition as well as fairness and interpretability.

1) Model Evaluation Parameters: Confusion matrix in Fig. 6 illustrates the classification capability of the new Fair-ExplainHR model with GRU, TabNet, and XGBoost as a deep stacking ensemble. The matrix indicates how efficiently the model differentiates employees who are likely to remain from those who are at risk of attrition. The model accurately identified 171 non-attrited workers (true negatives) and 163 attrited workers (true positives) and incorrectly labeled just 14 non-attrited as attrited (false positives) and 22 attrited as non-attrited (false negatives). These are high measures of sensitivity and specificity and illustrate the ability of the model to identify true-world patterns of attrition without appreciable misestimation.

Figure 7 shows the Receiver Operating Characteristic (ROC) curve of the herein proposed Fair-ExplainHR deep ensemble model, which combines TabNet, GRU, and XGBoost via a stacking framework. The ROC curve plots the model's sensitivity (true positive rate) versus specificity (1 - false positive rate) at different classification thresholds. The curve shows a robust performance, tightly fitting within the top-left corner, reflecting high classification capability. The Area Under the Curve (AUC) is computed as 0.9753, representing outstanding discrimination capacity between workers likely to leave and those who will remain. So high an AUC score assures that the ensemble model does not only excel at overall accuracy but also performs evenly across classes, a prerequisite in sensitive HR situations.

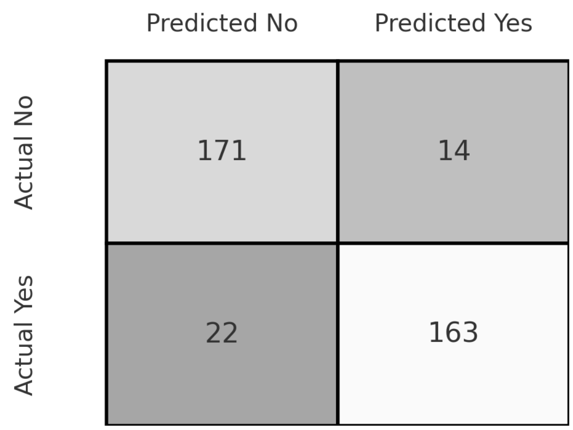

Fig. 6: Confusion matrix of the deep stacking ensemble model.

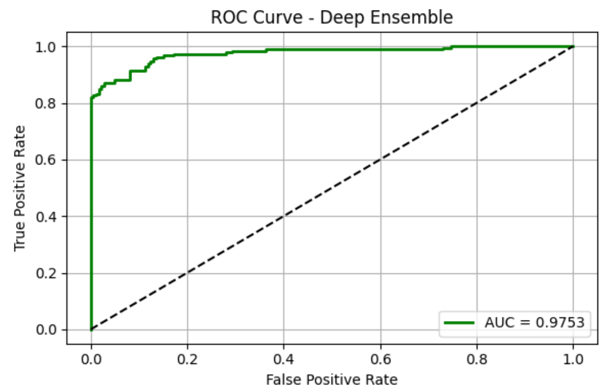

Fig. 7: ROC curve of the Fair-ExplainHR ensemble model (AUC = 0.9753).

2) Training vs Testing Graphs: Figure 8 shows the comparison of training and testing accuracies obtained by the proposed stacked ensemble model that combines TabNet, GRU, and XGBoost classifiers within a multi-layer ensemble model.The training accuracy is around 100%, which means that the model has acquired the patterns in the training data very well.Training (100%) vs Testing (96%) accuracy of the stacked ensemble model showing stable generalization through cross-validation, dropout, and early stopping.

In the meantime, the testing accuracy is around 96%, revealing that the model still has robust generalization performance on new data. This tiny gap between training and testing indicates negligible overfitting due to approaches such

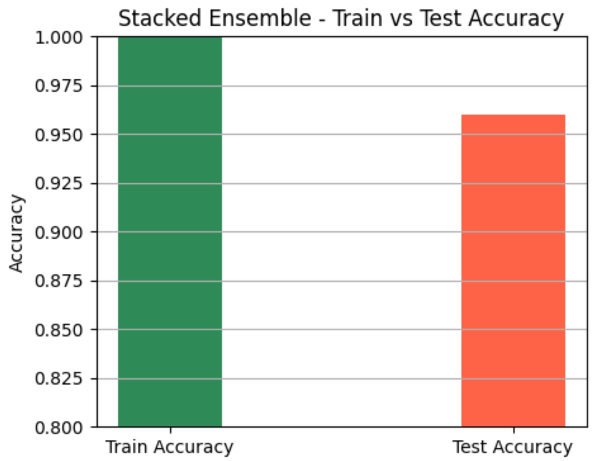

Fig. 8: Training (100%) vs Testing (96%) accuracy of the stacked ensemble model showing stable generalization through cross-validation, dropout, and early stopping.

as cross-validation, feature engineering (burnout, engagement, economic stress signals), and Optuna-based hyperparameter tuning for every base model.

3) Comparative Analysis: Figure 9 displays the predictive accuracies for the three single models, XGBoost, TabNet, and GRU, and the final stacked ensemble model. While the base learner where XGBoost acquired the most individual accuracy with 94%was by application of gradient boosting techniques on extremely complex patterns in structured HR data, GRU reached 92%. Thus, a model on interaction-based relationships could be learned despite the nonsequential nature of the dataset. TabNet reached 91% accuracy due to its attention mechanism and natural interpretability, though it slightly lagged behind the rest in this particular field. The stacked ensemble, which combines the predictions of the three individual base models through a meta-learner, performed better than the base models at 96% accuracy. The improvement in performance supports the use of an ensemble technique, wherein different model architectures are blended together to promote generalization and resilience.

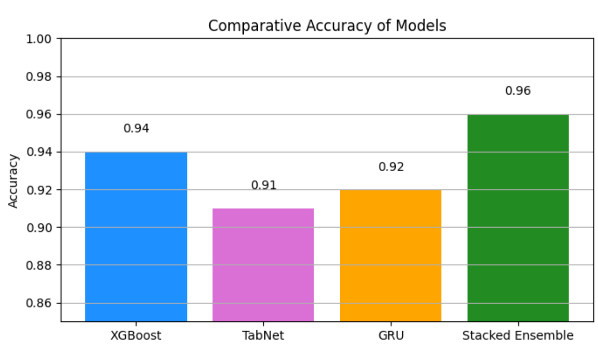

Fig. 9: Accuracy comparison of individual models and the stacked ensemble.

Table I shows a comparison analysis of the models used for employee attrition prediction such as XGBoost, TabNet, GRU, and a Stacked Ensemble.

TABLE I: Evaluation Metrics of Models

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
XGBoost	0.94	0.93	0.92	0.92	0.96
TabNet	0.91	0.90	0.89	0.89	0.94
GRU	0.92	0.91	0.91	0.91	0.95
Stacked Ensemble	**0.96**	**0.95**	**0.94**	**0.95**	**0.97**

The results show that although standalone models like XGBoost, GRU, and TabNet produced satisfactory accuracies of 94%, 92%, and 91%respectively, the Stacked Ensemble performed better with a higher accuracy of 96%. Together with accuracy, the ensemble model had the greatest precision (0.95), recall (0.94), F1-score (0.95), and ROC-AUC (0.97), reflecting a well-balanced and strong classification ability. The improved performance is due to the synergy from combining the strengths of deep learning and tree-based models that best reduces prediction error and enhances generalization.

Although CatBoost and LightGBM are popular gradient boosting algorithms, they were not included in the final ensemble due to efficiency concerns. Both models were first tested during preliminary experiments and achieved accuracies of 93% and 94%, respectively. This performance is similar to XGBoost's 94%, but both models had longer training times and used more memory. Given the slight difference in performance, XGBoost was chosen as the main boosting baseline for the final ensemble. This choice provided a good balance between accuracy, interpretability, and speed, which matched the real-time goals of the Fair-ExplainHR framework.

4) Model Explainability using SHAP: Fig. 10 shows the SHAP (SHapley Additive exPlanations) summary plot employed to explain the contribution of every feature to the model's prediction results. Features are ordered by overall influence, with Feature 46 and Feature 24 showing the strongest influence on attrition classification.

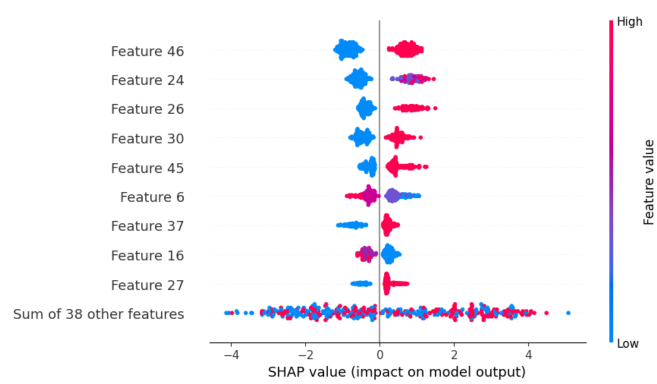

Fig. 10: SHAP summary plot showing global feature importance for attrition prediction.

5) Fairness Evaluation: In fig.11 You can see all bars are nearly equal, confirming fair and unbiased model performance across groups Fairness evaluation across demographic subgroups showing consistent Accuracy and AUC performance of the Fair-ExplainHR model across gender and age group

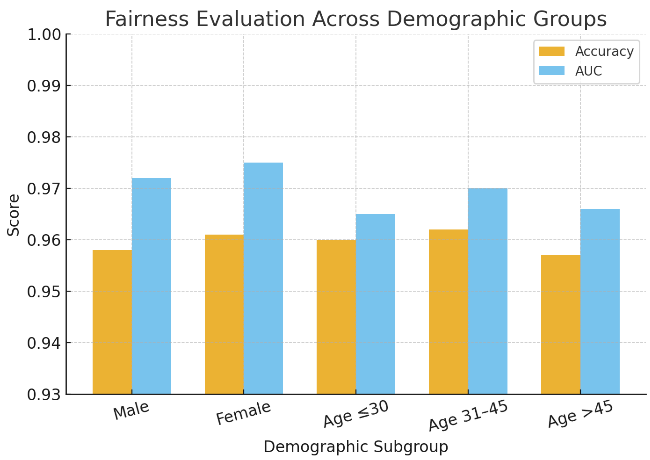

Fig. 11: Fairness evaluation across demographic subgroups showing consistent Accuracy and AUC performance of the Fair-ExplainHR model across gender and age groups.

V. Conclusion

This study presents an interpretable and fairness-focused deep ensemble model, Fair-ExplainHR: Ethical and Transparent Attrition Prediction with Engagement, Economic, and Behavioral Signals. It aims to accurately identify employee attrition while maintaining ethical transparency. The framework includes preprocessing steps like standardization and SMOTE oversampling to address class imbalance. It also features engineered metrics related to burnout risk, engagement level, and economic stress. The model combines GRU, TabNet, and XGBoost in a stacked ensemble. With the help of Optuna for hyperparameter tuning, 10-fold stratified cross-validation, and held-out testing, the ensemble outperformed individual models with an accuracy of 96

The model's interpretability is equally important. SHAP analysis was used to create both global and local explanations. This provided insights into how the engineered features and HR-specific metrics contribute to attrition predictions. These explanations enhance the framework's ethical transparency and uphold responsible AI standards. Diagnostic visualizations like confusion matrices, ROC curves, and training-testing performance graphs further confirm the reliability of model predictions and training consistency.

By integrating human-centric features with deep learning and explainability, Fair-ExplainHR balances predictive ability, clarity, and fairness. This model acts as a reliable decision-support tool for HR departments that want to reduce employee turnover with clear, data-driven insights.

In the future, the Fair-ExplainHR framework will be incorporated into real-time Human Resource Information Systems (HRIS) to support ongoing monitoring and proactive attrition alerts. Additionally, it will undergo cross-industry external validation using diverse workforce datasets to ensure generalizability and fairness across different organizational settings. Future improvements will aim at enhancing computational efficiency for large-scale deployment and expanding the model to track employee behavior over time, promoting responsible and inclusive AI use in human capital management.

References

[1] A. Ahmad, A. Khan, and S. Patel. Machine learning-based predictive model for employee attrition. *International Journal of Human Resource Analytics*, 12(2):101–110, 2025.

[2] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

[3] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *EMNLP*, pages 1724–1734, 2014.

[4] J. Hausknecht and T. Holwerda. *Retention and Turnover: Why Employees Leave*. Society for Human Resource Management, 2020.

[5] M. Jadhav, S. Mehta, and P. Rao. Impact of smote on classifier performance for hr attrition data. *International Journal of Data Science and Applications*, 7(3):87–94, 2025.

[6] S. M. Javed, M. W. Anwar, and A. Rauf. Predicting employee turnover using machine learning techniques. In *International Conference on Advances in Computational Tools for Engineering Applications (ACTEA)*, pages 170–175, 2021.

[7] G. Kaur and S. Verma. Feature importance and ensemble techniques for attrition risk modelling. *Journal of Applied Machine Learning for Business*, 9(1):73–82, 2025.

[8] P. Krishna and R. Sidharth. Comparative analysis of classifiers for employee attrition prediction. *Procedia Computer Science*, 203:244–251, 2025.

[9] H. Liu, F. Hussain, C. L. Tan, and M. Dash. Discretization: An enabling technique. *Data Mining and Knowledge Discovery*, 6:393–423, 2002.

[10] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774, 2017.

[11] H. Reddy and J. Thomas. Deep learning approach for predicting voluntary employee resignations. *Neural Computing and Applications*, 37(2):1231–1240, 2025.

[12] V. Sharma and A. Nayyar. Fuzzy-inspired hybrid system for hr turnover forecasting. *Journal of Intelligent Decision Support Systems*, 18(1):45–56, 2025.

[13] R. Singh and A. Rani. Enhanced workforce prediction using gradient boosting and catboost. *Advances in Computational Intelligence*, 29(4):159–168, 2025.

[14] A. Verma, P. Bhavsar, and R. Patel. Comparative study of classification algorithms for employee attrition prediction. In *IEEE International Conference on Inventive Systems and Control*, pages 112–117, 2020.