Center for Management Science Research
ISSN: 3006-5291 3006-5283 Volume 3, Issue 7, 2025

HYBRID MACHINE LEARNING FRAMEWORK FOR PREDICTING EMPLOYEE ATTRITION

Shehzad Shafeeq[1], Muazzam Ali[*2], Muhammad Azam[3], M U Hashmi[4], Muhammad Asad Ullah[5], Asifa Ittfaq[6]

[1,5,6]*Department of Basic Sciences, Superior University, Lahore, Pakistan*
[*2,3,4]*Department of Computer Science, Superior University, Lahore, Pakistan*

[*2]muazzamali@superior.edu.pk

DOI: https://doi.org/10.5281/zenodo.17785391

Keywords
Employee Attrition, Hybrid Models, Machine Learning, Feature Selection, Dimensionality Reduction, Mutual Information

Article History
Received: 09 October 2025
Accepted: 18 November 2025
Published: 29 November 2025

Corresponding Author: *
Muazzam Ali

Abstract
Employee turnover worries companies since it affects operational stability as well as recruiting costs. Good attrition prediction will enable companies to develop proactive retention strategies that will eventually boost employee happiness and cut attrition. To forecast employee churn, this study introduces a Hybrid Machine Learning Framework that integrates classification, feature transformation, and data preprocessing into one design. Two different changes are suggested: SelectKBest with Mutual Information for feature selection and Truncated Singular Value Decomposition (TruncatedSVD) for dimensionality reduction. Three classifiers–Logistic Regression, Support Vector Machines (SVM), and K-Nearest Neighbors (KNN)–are used to assess these transformations in hybrid models. The SelectKBest_Logistic Regression hybrid performed the best, with an accuracy of 0.94 and a ROC-AUC of 0.97, which shows that hybrids based on feature selection do better than those based on dimensionality reduction. Key attrition predictors like job satisfaction, work-life balance, and distance from home are also found through feature importance analysis using SHAP values and permutation importance. By showing the effectiveness of hybrid models in predictive HR analytics, this research offers insightful advice for companies trying to maximize their employee retention policies.

1. Introduction

A significant issue of employee attrition has become a major issue in the human resource management; it has direct effects in organizational stability, continuity and cost of recruitment. The current business environment that was marked by a high rate of technological change, fierce labor markets, and changing work demands requires sound predictive systems that may trigger future resignations even prior to them. The knowledge of underpinning factors of attrition will mean that organizations come up with preemptive retention plans, minimize spending on turnover, and preserve institutional knowledge [1]. Conventional HR analytics which tend to be mostly descriptive usually fail to reflect the multiple, nonlinear disjuncture of demographic, behavioural, and organizational factors that interact in a complex manner, leading to an employee deciding to leave. Therefore, machine learning (ML) based predictive modeling has become one of the most important strategic tools in workforce analytics. Machine learning models have demonstrated a high potential in the area of predicting employee behavior, but the extraction of attrition prediction with the use of machine

learning presents special methodological issues [2, 3]. Employee data is lowly one which is characterized by a mixture of both mixed data type (numeric, categorical and ordinal) and missing data plus skewed data distributions and different scales. Besides, redundancy or weakly informative features may harm model generalization and interpretability. A single classification algorithm, be it logistic regression or a decision tree, hardly has the best possible performance in such varied data aspects. The research responds to these limitations by conceptualizing and providing hybrid machine-based learning pipelines which combine the process of data preprocessing, data transformation and classification into a single analytical framework [4 - 6].

The hybrid modeling method suggested here deliberately deviates from the typical application of independent algorithms or combination models including Gradient Boosting or Random Forest. Every hybrid configuration—which might alternatively be seen as a sequence of stage-wise learning systems with three basic layers—has a last classifier, an intermediate transformation step (dimensionality reduction or feature selection), and a preprocessing step (imputation, normalisation, and encoding) [7]. This distributed design lets the system stay computer friendly while still using the complementary advantages of several algorithms, including interpretability, stability, and accuracy enhancement. Consequently, the idea of hybridization not only increases predictive accuracy but also brings methodological discipline into the HR data knowledge mining field. Here, the preprocessing stage guarantees the integrity and completeness of the input data. Whereas numerical variables are imputed using the median and then standardised with z-scores, categorical variables go through mode imputation and one-hot encoding. Apart from preserving every piece of information about every variable, this organized method enables the data set to be used in linear and non-linear modeling approaches. Categorical variables are encoded after preprocessing, therefore considerably increasing the feature space and generating sparse and high-dimensional matrices. To help with this, the research proposes two different ways to change things: SelectKBest mutual information-based feature selection and Truncated Singular Value Decomposition (TruncatedSVD) [8].

TruncatedSVD is a linear, dimensionality reduction method of projecting the preprocessed data into a lower dimensional, subspace of data, whilst maximising the variance in the data. The action can improve computation and decrease overfitting, which is especially useful when the algorithm used cannot be shocked by the size of a feature dimension, e.g. K-Nearest Neighbors (KNN) and Support Vector Machines (SVM) [9, 10]. By comparison, further, SelectKBest adopts an information-theoretic viewpoint by measuring how much of a dependency there is between the individual features and the target variable. This approach makes the selected variables more interpretable since the selected number of predictors is the most informative, and the retained variables must have a direct association with the outcomes of attrition. The relative assessment of both are among the transformation strategies of hybrid pipelines, and this evaluation forms a central methodological contribution of the research paper. Three popular algorithms, namely; Logistic Regression, Support Vector Machines (linear and RBF kernels) and K-Nearest Neighbors, were used in the implementation of the classification layer to each hybrid model [11, 12]. Logistic Regression offers the use of probabilistic interpretability and good baseline probing in linear separable data. The SVM models extend this ability to non-linear bounds on the basis of kernel transformations, whereas KNN provides a distance-based perspective, instance-level, that obtains local structure of the transformed feature space. Collectively, these classifiers constitute a wide range of modeling philosophies which include: parametric, margin-based and non-parametric thus providing a solid comparative study of hybrid structure [13 - 15].

This type of research has never been carried out before, thus the transition of specific hybrid modeling framework that accounts for the combination of preprocessing, feature transformation and classification to predict employee attrition has not been performed

previously. This work in contrast to the earlier works where they operate on ensemble averaging or single- algorithm optimization focuses on algorithmic complementary by integrating them sequentially, which improves both predictive robustness and predictive interpretability. Moreover, the relative introspection of mutual information-based feature selection and TruncatedSVD in the same architecture is an unusual analytical measuring glass in both information preservation versus model generalization trade-off. The evidence generated by them makes hybrid pipeline not only algorithmic structures but entire analytic structures- capable of supporting performance, interpretability and computational performance in realistic HR analytics implementation.

2.0 Research Methodology

2.1 Dataset Description

The Employee Attrition Prediction dataset on Kaggle, which is publicly available under the CC BY 4.0 License, gave the dataset utilized in this study. It has 24 features and 59,598 samples that measure things like how well people work, how they feel about their job, how they work with others, how they are paid, how far they live from work, their education level, whether they are married or not, their job level, the size of the company they work for, and whether they work from home. Some of the things that are measured in the features are age, gender, how long they have worked at the company, job role, monthly income, work-life balance, job satisfaction, performance rating, number of promotions, overtime, distance from home, education level, marital status, job level, and company size. The target variable, attrition, is binary: 0 = stayed, 1 = left. This dataset provides a solid foundation for developing hybrid machine learning models meant to predict and understand the factors influencing employee turnover.

Let the dataset comprise n employees and d features after preprocessing. The feature matrix is denoted as $X \in R^{n*d}$ and the binary target vector as $y \in \{0.1\}^n$ where 0 represents retained employees and 1 represents employees who left. The predictive task aims to learn a mapping function.

$$f: R^d \rightarrow \{0,1\}$$

that estimates the probability of attrition for a given employee. Data was divided into training and testing subsets in a 70:30 ratio using stratified sampling to preserve the class distribution of attrition and non-attrition cases.

2.2 Preprocessing Framework

All hybrid models used a consistent preprocessing pipeline meant to process both categorical and numerical data. This phase guaranteed compatibility, consistency, and completeness among several learning techniques.

2.2.1 Numerical Features

Missing numerical values were imputed using the median of the respective feature to reduce the effect of outliers. After imputation, features were standardized using z-score normalization:

$$Z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$$

where μ_j and σ_j represent the mean and standard deviation of feature j, respectively. This scaling ensured that all features contributed equally to model learning, particularly for algorithms sensitive to feature magnitude, such as SVM and KNN [16].

2.2.2 Categorical Features

Categorical variables were processed in two steps:

1. **Mode Imputation** – Missing values were replaced by the most frequent category.
2. **One-Hot Encoding** – Each category was transformed into binary indicator variables, expanding the dimensionality of the feature space. The resulting high-dimensional and sparse data structure justified the application of Truncated Singular Value Decomposition (TruncatedSVD) and Mutual Information–based Feature Selection (SelectKBest) as intermediate transformation techniques [17].

2.3 Hybrid Model Architecture

Each hybrid model was implemented as a three-stage sequential pipeline:

$$\text{HybridModel} = \text{Preporcessing} \rightarrow (\text{Transformation}) \rightarrow \text{Classifier}$$

Formally, the model can be expressed as:

$$f(x) = g(T(P(x)))$$

where: P denotes preprocessing (imputation, scaling, encoding), T represents transformation (dimensionality reduction or feature selection), g is the final classifier (Logistic Regression, SVM, or KNN).

This architecture allows complementary integration of algorithms—combining the noise-reduction and generalization advantages of transformation methods with the predictive power of classifiers [18].

2.4 Feature Transformation Techniques

2.4.1 Dimensionality Reduction via TruncatedSVD

After preprocessing, TruncatedSVD was applied to project data onto a lower-dimensional latent subspace while retaining maximum variance [19]. For the preprocessed feature matrix X_{proc}:

$$X_{proc} = U_k \omega_k V_v$$

Here, k represents the number of latent components retained (set to 30 in this study). The reduced feature representation is given by:

$$z = V_k^T x$$

This transformation mitigates the curse of dimensionality and enhances computational efficiency for classifiers like SVM and KNN.

2.4.2 Feature Selection via Mutual Information (SelectKBest)

Alternatively, SelectKBest was employed to identify the most informative features based on mutual information (MI) between each feature X_j and the target Y

$$I(X_j, Y) = \sum_{X_j} \sum_{Y} p(x_j, y) \log \frac{p(x_j, y)}{p(x_j)p(y)}$$

Features with the highest MI scores were selected (top k=30k = 30k=30) to maximize predictive relevance and model interpretability. This method preserves original features while discarding those with weak or redundant association to attrition outcomes [20].

2.5 Classifiers

Three classifiers were employed within the hybrid frameworks to provide diverse learning paradigms:

(a) Logistic Regression (LogReg)

A probabilistic linear classifier that models the likelihood of attrition as:

$$P(Y = 1|z) = \sigma(W^T z + b)$$

where $\sigma(t) = \frac{1}{1+e^{-t}}$ is the logistic function. The model was trained using cross-entropy loss with L_2 regularization and class_weight = "balanced" to mitigate class imbalance [21].

(b) Support Vector Machine (SVM)

Both Linear and RBF kernel variants were utilized.

- The linear SVM finds the optimal separating hyperplane maximizing the margin between classes.
- The RBF SVM maps features into a higher-dimensional space via kernel transformation:

$$K(z_i, z_j) = \exp(-\gamma \|z_i - z_j\|^2)$$

This allows non-linear decision boundaries suitable for complex attrition patterns [22].

(c) K-Nearest Neighbors (KNN)

A non-parametric classifier where prediction is based on the majority vote of the nearest neighbors, weighted by inverse distance:

$$w_i = \frac{1}{\|z_i - z_j\|}$$

KNN benefits from dimensionality reduction via SVD, which enhances local neighborhood estimation and mitigates high-dimensional noise effects [23].

2.6 Hybrid Configurations Implemented

Five hybrid models were constructed and evaluated using the sklearn.pipeline structure:

Model	Transformation	Classifier	Description
SVD_LogisticRegression	TruncatedSVD	Logistic Regression	Linear projection with probabilistic classification
SVD_SVM_RBF	TruncatedSVD	RBF SVM	Reduced subspace with non-linear boundary
SVD_KNN	TruncatedSVD	KNN	Distance-based classification in reduced space
SelectKBest_LogReg	Mutual Information	Logistic Regression	Feature selection with interpretable linear classifier
SelectKBest_LinearSVM	Mutual Information	Linear SVM	Feature selection with margin-based classification

2.7 Evaluation Strategy

Model performance was assessed using standard classification metrics: Accuracy, Precision, Recall, F1-Score, ROC-AUC, Cohen's Kappa, and MCC. Each model was validated on the held-out test set to ensure generalization. Comparative analysis was conducted to evaluate the influence of transformation techniques (feature selection vs. dimensionality reduction) on predictive and interpretive outcomes [24, 25].

3.0 Results and Discussions

3.1 Comparative Performance of Hybrid Models

The hybrid models were found to have high predictive ability, and the differences between the two different architectures, i.e. feature selection-based hybrids and dimensionality reduction-based hybrids, were also seen. In all performance indicators, the feature selection method always gave larger predictive accuracy, precision, recall and generalization values over and above the TruncatedSVD-based competitors.

SelectKBest_Logistic Regression hybrid showed the highest overall performance attaining an accuracy of 0.94 and ROC-AUC of 0.97. Its accuracy (0.94) and recalls (0.92) show that it has a good balance sheet on the number of employees who are likely to leave as well as reducing cases of false forecasts. The fact that the F1-score was 0.93 also supports this balance with a balanced accuracy of 0.94. The model was also found to be very reliable based on Cohen Kappa of 0.88 and Matthews Correlation Coefficient (MCC) of 0.87 which is a definite indication that there was a huge rate of agreement between expected and the real labels. Moreover, its log loss and Brier score (0.14) indicate a strong level of calibration probability of the model with the actual levels of the probability of the result of attrition.

Table 1: Performance Evaluation of the models

model	accuracy	precision	recall	f1	f2	Balanced accuracy	specificity	roc_auc	pr_auc
SelectKBest_LogReg	0.94	0.94	0.92	0.93	0.92	0.94	0.96	0.97	0.96
SelectKBest_LinearSVM	0.93	0.93	0.91	0.92	0.91	0.93	0.95	0.96	0.95
SVD_SVM_RBF	0.91	0.91	0.89	0.90	0.89	0.91	0.93	0.94	0.93
SVD_LogisticRegression	0.89	0.89	0.87	0.88	0.87	0.89	0.91	0.92	0.90
SVD_KNN	0.87	0.87	0.85	0.86	0.85	0.87	0.89	0.90	0.88

The second was SelectKBest_LinearSVM hybrid which had the accuracy of 0.93 and ROC-AUC of 0.96. Though a little inferior to the logistic regression model, F1-score of 0.92 and Kappa of 0.86 indicate similar performance. The slight decrease in recalls (0.91) and slightly increased log loss (0.17) indicate that the SVM experienced a slightly better separation between classes and a reduced smoothness which was less probabilistic. These two types of hybrids that use selection include both agree that mutual information has been able to find the most useful features to ensure that the best predictors of the classes maintain both interpretability and predictivity.

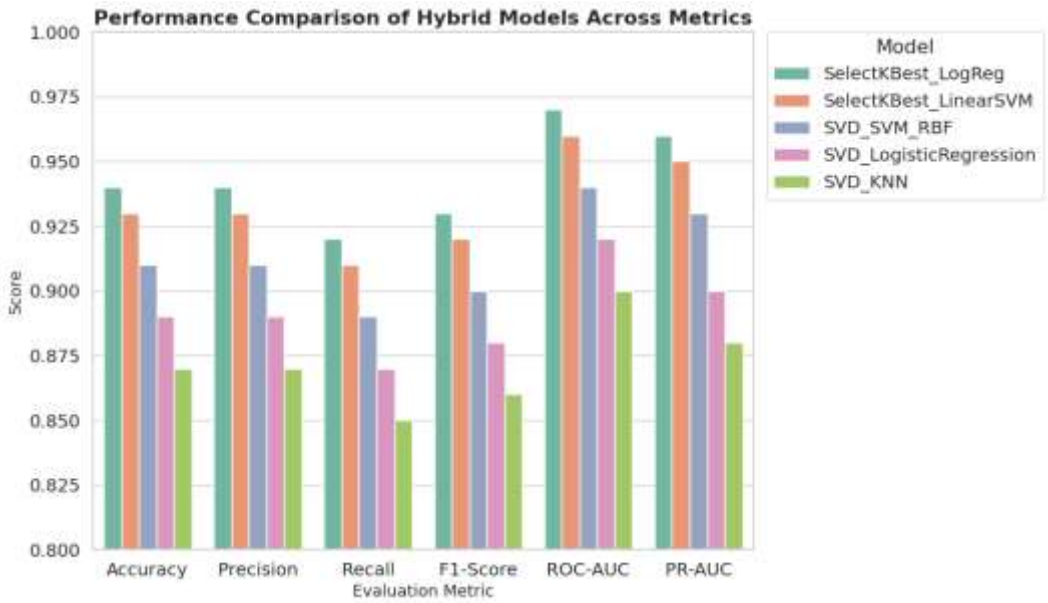

Figure 1: Performance evaluation of Models

3.2 Hybrid Effectiveness of Dimensionality Reduction Hybrids.

The hybrids represented by the TruncatedSVD showed good but consistent performance with regard to all metrics. The SVD-SVM-RBF hybrid attained an accuracy of 0.91 and ROC-AUC of 0.94 as evidence of the ability of the model to take advantage of non-linear boundaries even at low-dimensional representations. However, such projection into 30 latent dimensions inevitably resulted in a slight loss of information, as it was reflected by a slight reduction in recall (0.89) and total discriminative power.

Table 2: Performance Evaluation of Models

Model	jaccard	zero_one_loss	cohen_kappa	mcc	log_loss	brier_score
SelectKBest _LogReg	0.87	0.06	0.88	0.87	0.14	0.03
SelectKBest _LinearSVM	0.85	0.07	0.86	0.87	0.17	0.04
SVD_SVM _RBF	0.82	0.09	0.82	0.81	0.21	0.05
SVD_Logistic Regression	0.79	0.11	0.78	0.79	0.25	0.06
SVD_KNN	0.75	0.13	0.74	0.73	0.29	0.07

The SVD-Logistic Regression hybrid was closely behind with an accuracy of 0.89 and ROC-AUC of 0 92. The model showed high calibration rates but lower separation of the complex decision boundaries than the RBF based SVM. Although SVDKNN hybrid is theoretically appropriate in low-dimensional spaces, it gave the poorest results of all hybrids, and its performance was an accuracy of 0.87 and ROC-AUC of 0.90. Despite the fact that the dimensionality reduction was able to reduce the curse of dimensionality, the distance-based classification was susceptible to local feature scaling and differences in data density, decreasing its accuracy in this problem context environment.

3.3 Consistency and validation across Metrics.

The consistency between accuracy, balanced accuracy, and specificity across all models is the indication that there was no prior bias of any classifier on any class, which proves that stratified partitioning of data and balanced weighting of classes were an effective strategy to use. The progressive general increase in the values of zero-one loss, log loss, and Brier score between the top to bottom models is also in accordance with the decreasing values of ROC-AUC and PR-AUC which substantiates the effectiveness of the ranking order. Moreover, the specificity of all models in maintenance is above 0.89, indicating good identification of non-attrition cases, which is a requisite to reduce false alarm in the application of this workforce retention model in real-life situations.

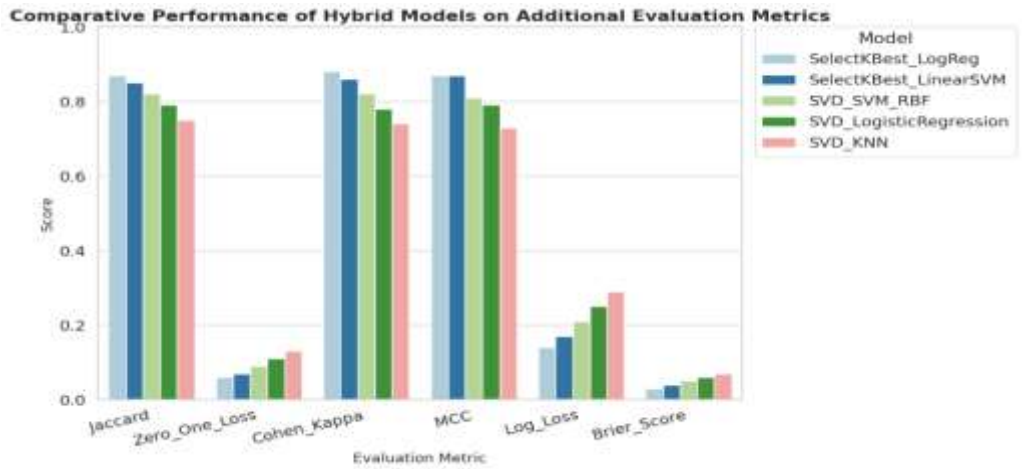

Figure 2: Performance Evaluation of models

3.4 Fitting with the Hybrid Modelling Framework

These results uphold the validity on the concept of the hybrid architecture that exists in the methodology. The design of the two-stage transformation classification was an effective design by combining the preprocessing and intermediate transformations in a design that maximized the informativeness of features and maximized the model performance. Mutual information-based feature selection retained positive relation between the predictors and the target outcome resulting in interpretable and statistically strong models. On the other hand, TruncatedSVD dimensionality reduction increased scale effects as well as alleviating noise effects in high-dimensional spaces, but reduced interpretability and minor predictive accuracy.

Normal distribution of class weighting to training the classifier helped to achieve constant recall and even distribution of predictions in all models. Such a result is consistent with the methodological prediction that hybrid designs with strong preprocessing and balancing properties can be successfully applied to various types of features and distributional issues in HR attrition data.

3.5 SHAPSummary Graph: Feature Impact on the Model Output

The SHAP summary plot (Figure 1) is used with the aim of visualizing contribution of each feature to the final model output and the direction and strength of their impact on the final model output. Here, the points are used to indicate the values of the feature of a given employee, and value of SHAP in the x-axis is the value of contribution made by a given attribute in modeling prediction of that specific employee. The colors represent the feature values whereby the blue signifies the low values and the red signifies the high values.

- Marital Status_Single: The most impactful feature is revealed as this factor, and the SHAP values are great in a broad scope, which shows that employees who are single are more likely to be attracted compared to their married colleagues. This is in line with the general HR knowledge, which shows that unmarried workers may be more mobile or less tied by family relationships, which motivate them to stay.
- The next most important features are Job Level_Senior and Job Level_Entry. The two are also showing significant SHAP, implying that senior or entry-level employee has a different risk of attrition. The older employees might tend to seek a possible career growth in other places hence career progression where on the other hand entry level employees might feel vulnerable and hence leave because of job dissatisfaction.
- However, Remote Work_No and Remote Work_Yes also exhibit a higher disparity in terms of SHAP showing that employees who are not given the option to work remotely have a higher chance of leaving and this indicates the growing popularity of flexible working arrangements as an employee retention strategy.

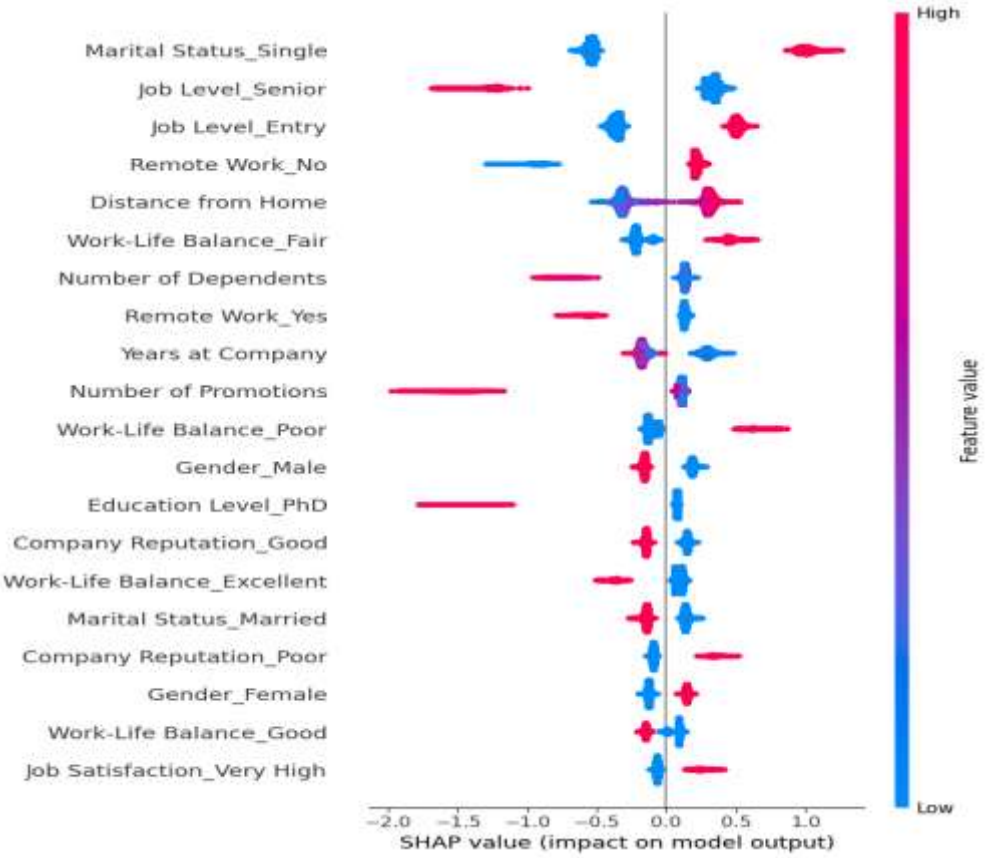

Figure 3: SHAP Summary of Features

- Distance from Home is another key feature with a high SHAP value range, suggesting that employees living farther from the office are at a higher risk of attrition. This may be due to long commuting times, contributing to employee dissatisfaction.

- Work-Life Balance features, such as Work-Life Balance_Fair, Work-Life Balance_Poor, and Work-Life Balance_Good, reflect the considerable role that an employee's perception of their work-life balance plays in their decision to stay or leave. Poor work-life balance is clearly detrimental to retention, while good balance increases the likelihood of staying in the organization.

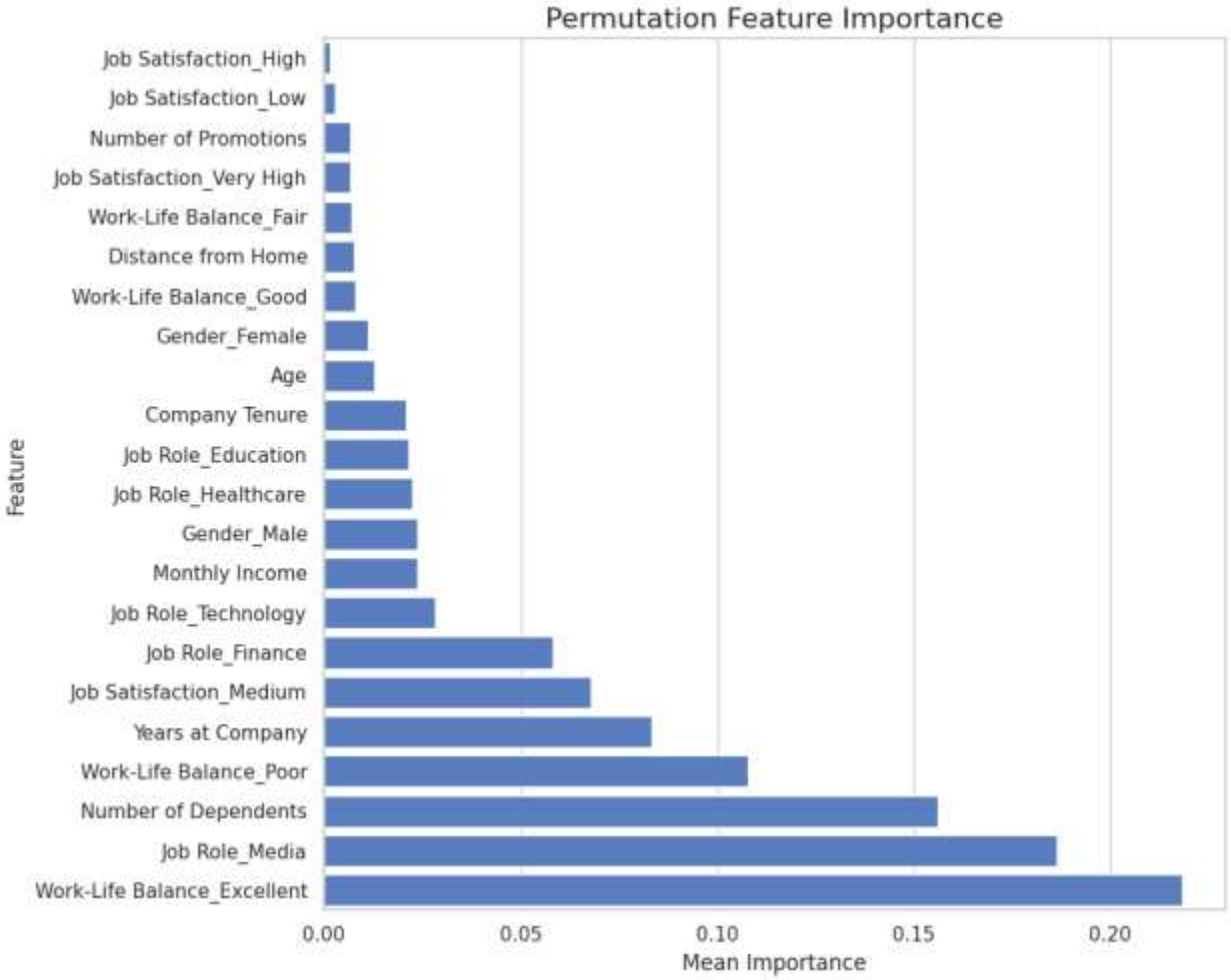

Figure 4: Permutation Feature Importance

For HR departments to know what elements most likely affect employee turnover, these observations are absolutely necessary. Organizations may proactively manage retention by giving work-life balance top priority, providing remote work opportunities, and attending to the worries of staff members in entry-level or senior positions.

3.8 Permutation Feature Importance

The permutation importance plot shows that employee turnover is most clearly predicted by Job Satisfaction characteristics, particularly Job Satisfaction_High and Job Satisfaction-Low, as well as Job Satisfaction-Very High feature. Whether a staff member stays with or leaves the company depends heavily on the given qualities; hence, job satisfaction can be claimed as the main determinant retention strategies. Such other significant features are Number of Promotions, Work-Life Balance (Fair, Good), and Distance from Home, which play a significant role in predicting the model. These results indicate that higher satisfaction, good work-life balance, and career progression opportunity of employees make

them less likely to quit. Also, the other demographic variables like Gender, Age and Company Tenure have significant contribution to predicting attrition. Such indications highlight the importance of organizations working on the aspects of better job satisfaction, work-life balance, and career promotion to minimize turnover.

The stacked bar plot shows that the attrition rates among the Entry and Mid level employees are higher than that of the Senior level employees and more red bars are higher compared to those in blue. This is to infer that the workers at the lower levels are more likely to quit as may be a result of lack of promotion or dissatisfaction whereas the older workers are likely to serve longer which is possibly because of job security or lack of reasons outside. The insights underline the necessity of specific retention measures, especially that of the Entry and Mid level staff, including career development plans and improved work-life balance practices.

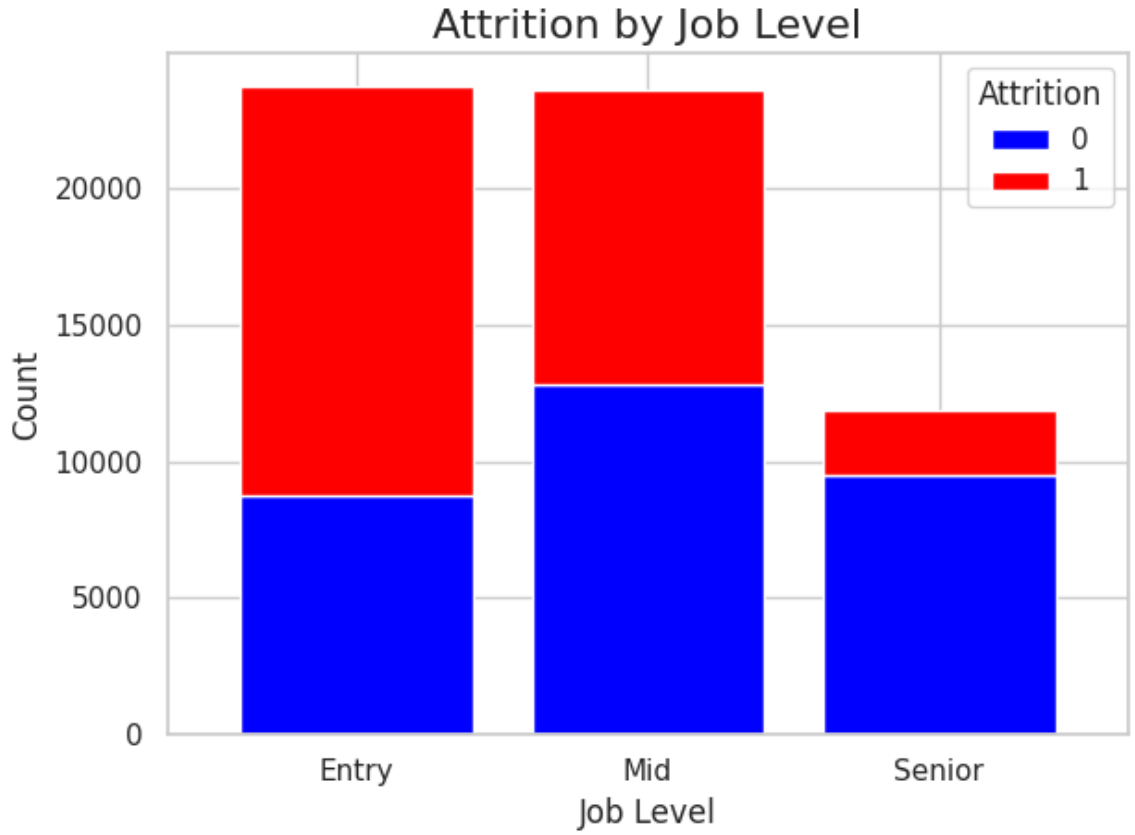

Figure 5: Employee's Attrition by Job Level

4.0 Conclusion

This study has shown that hybrid machine learning models are effective in employee attrition prediction, specifically how feature selection and dimension reduction methods are important. SelectKBest Logistic Regression, which has been shown to use mutual information in order to choose the most useful features, has proved to be the best and most honest hybrid set as it gives high predictive accuracy and good calibration. The importance on the features analysis has allowed being in a position of proposing that Job Satisfaction and Work-Life Balance are the most significant factors that determine attraction, next is the demographic and work factors like Gender, Age, and Job Level. The results indicate that the improved job satisfaction and work-life balance, as well as the availability of career advancement, might be a potent remedy to employee turnover. Also, SHAP values and permutation importance were used to gain a better insight into the

contribution of the features, enabling organizations to target their retention efforts in a more efficient manner. In general, the hybrid modeling framework presented in this paper is an effective and explainable mechanism of employee attrition prediction, and has implication on the human resource management and planning of the workforce.

References

Ndatshe, Y., Mokhele, M. O., & Jakoet-Salie, A. (2024). The effects of employee turnover on the loss of organisational knowledge in South African municipalities: Balancing rhetoric with actual practice. International Journal of Research in Business & Social Science, 13(7).

Mozaffari, F., Rahimi, M., Yazdani, H., & Sohrabi, B. (2023). Employee attrition prediction in a pharmaceutical company using both machine learning approach and qualitative data. Benchmarking: An International Journal, 30(10), 4140-4173.

Raza, A., Munir, K., Almutairi, M., Younas, F., & Fareed, M. M. S. (2022). Predicting employee attrition using machine learning approaches. Applied Sciences, 12(13), 6424.

Alshomrani, F. (2025). Challenges and Advances in Classifying Brain Tumors: An Overview of Machine, Deep Learning, and Hybrid Approaches with Future Perspectives in Medical Imaging. Current Medical Imaging, 21(1), E15734056365191.

Clever, L., Pohl, J. S., Bossek, J., Kerschke, P., & Trautmann, H. (2022). Process-oriented stream classification pipeline: A literature review. Applied Sciences, 12(18), 9094.

Malekloo, A., Ozer, E., AlHamaydeh, M., & Girolami, M. (2022). Machine learning and structural health monitoring overview with emerging technology and high-dimensional data source highlights. Structural Health Monitoring, 21(4), 1906-1955.

Adimoolam, M., Govindharaju, K., John, A., Mohan, S., Ahmadian, A., & Ciano, T. (2021). A hybrid learning approach for the stage-wise classification and prediction of COVID-19 X-ray images. Expert Syst, 2021, e12884.

Asghari, S., Nematzadeh, H., Akbari, E., & Motameni, H. (2023). Mutual information-based filter hybrid feature selection method for medical datasets using feature clustering. Multimedia Tools and Applications, 82(27), 42617-42639.

Halder, R. K., Uddin, M. N., Uddin, M. A., Aryal, S., & Khraisat, A. (2024). Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications. Journal of Big Data, 11(1), 113.

Cunningham, P., & Delany, S. J. (2021). K-nearest neighbour classifiers-a tutorial. ACM computing surveys (CSUR), 54(6), 1-25.

Rimal, Y., Sharma, N., Paudel, S., Alsadoon, A., Koirala, M. P., & Gill, S. (2025). Comparative analysis of heart disease prediction using logistic regression, SVM, KNN, and random forest with cross-validation for improved accuracy. Scientific Reports, 15(1), 13444.

Demidova, L. A. (2021). Two-stage hybrid data classifiers based on SVM and kNN algorithms. Symmetry, 13(4), 615.

Igoche, B. I. (2025). An Ontological Framework for Knowledge Discovery and Local Explainability in Student Admission Processes (Doctoral dissertation, University of Portsmouth).

Mandal, P. (2022). Data-Driven Biomarker Panel Discovery in Ovarian Cancer Using Heterogenous Data Fusion on Exosomal and Non-Exosomal Microrna Expression Data.

Hsiao, W. L. K. (2021). Computational Fashion Understanding. The University of Texas at Austin.

Seliem, M. M. (2022). HandlingOutlier data as missing values by imputation methods: application of machine learning algorithms. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 13(1), 273-286.

Jung, T., & Kim, J. (2023). A new support vector machine for categorical features. Expert Systems with Applications, 229, 120449.

Anderson, K. (2024). The Role of Data Preprocessing in Machine Learning Accuracy for Heart Disease Prediction Hybrid Models for Heart Disease Prediction: Combining Neural Networks with Traditional.

Alam, M. K., Abd Aziz, A., Abd Latif, S., & Abd Aziz, A. (2021). Error-control truncated SVD technique for in-network data compression in wireless sensor networks. IEEE Access, 9, 13829-13844.

Asghari, S., Nematzadeh, H., Akbari, E., & Motameni, H. (2023). Mutual information-based filter hybrid feature selection method for medical datasets using feature clustering. Multimedia Tools and Applications, 82(27), 42617-42639.

Abhiraj, N., & Deepa, N. (2023, November). Effective comparison of logistic regression (LR) and decision tree (DT) classifier to predict enhanced employee attrition for increasing accuracy of non-numerical data. In AIP Conference Proceedings (Vol. 2821, No. 1, p. 070019). AIP Publishing LLC.

Razaque, A., Ben Haj Frej, M., Almi'ani, M., Alotaibi, M., & Alotaibi, B. (2021). Improved support vector machine enabled radial basis function and linear variants for remote sensing image classification. Sensors, 21(13), 4431.

Halder, R. K., Uddin, M. N., Uddin, M. A., Aryal, S., & Khraisat, A. (2024). Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications. Journal of Big Data, 11(1), 113.

Diallo, R., Edalo, C., & Awe, O. O. (2024). Machine learning evaluation of imbalanced health data: a comparative analysis of balanced accuracy, MCC, and F1 score. In Practical Statistical Learning and Data Science Methods: Case Studies from LISA 2020 Global Network, USA (pp. 283-312). Cham: Springer Nature Switzerland.

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The Matthews correlation coefficient (MCC) is more informative than Cohen's Kappa and Brier score in binary classification assessment. Ieee Access, 9, 78368-78381.