Employee Attrition Prediction Using Machine Learning

K. R. Rohith Kumar[1],S Rajasekhar[2], Vuppu Sri Sucharitha[3],
Bommisetti Nethan[4], G Narasimha Reddy[5], Naruboyina Vijay Kumar[6]
[1,2]*Assistant Professor, Department of Artificial Intelligence and Machine Learning,*
Annamacharya Institute of Technology & Sciences, Tirupati, India
[3,4,5,6]*Student, Department of Artificial Intelligence and Machine Learning,*
Annamacharya Institute of Technology & Sciences, Tirupati, India

***Abstract*—The increasing adoption of machine learning techniques by organizational decision-makers has encouraged researchers to investigate their applicability in addressing critical workforce challenges. Employee attrition, particularly the loss of skilled and experienced personnel, remains a significant concern for modern organizations. This study examines the effectiveness of machine learning models in predicting employee attrition using a synthetic dataset provided by IBM Watson. Three experimental scenarios were designed to evaluate model performance. In the first scenario, machine learning algorithms Support Vector Machine (with multiple kernel functions), Random Forest, and K-Nearest Neighbors were trained on the original class-imbalanced dataset. The second scenario addressed class imbalance through the Adaptive Synthetic Sampling (ADASYN) technique, followed by retraining the same models on the balanced data. The third scenario applied manual undersampling to achieve class balance. Experimental results indicate that the ADASYN-balanced dataset combined with the K-Nearest Neighbors algorithm (K = 3) produced the best performance, achieving an F1-score . Additionally, feature selection techniques integrated with the Random Forest model yielded an F1-score while reducing the feature set from 29 to 12 attributes. The findings demonstrate the importance of data balancing and feature optimization in enhancing employee attrition prediction.**

***Index Terms*—hand gesture recognition, deep learning, convolutional neural networks, CNN, human-computer interaction, sign language, image processing, real-time recognition, gesture classification, assistive technology**

I. INTRODUCTION

1.1 Background And Motivation

The loss of employees regardless of the following reasons can be referred to as employee attrition; personal reasons, low job satisfaction, low salary, and a poor business environment. The employee attrition may be classified into two types; voluntary and involuntary attrition. Involuntary attrition takes place when the employer of employees sacks them due to various reasons like poor employee performance or business need. Under voluntary attrition, however, high performance employees choose to quit the company voluntarily even when the company is making efforts to keep them. Early retirement or offers of employment in other companies, etc., may lead to voluntary attrition. Despite the fact that those companies, where executives have understood the significance of their staff, often invest into their personnel by giving high-quality training and offering employees a wonderful work climate, they also experience the same issue of voluntary turnover and losing their brightest employees. The other problem, the replacement hiring cost, is expensive to the company in terms of interviewing, hiring and training. Anticipating the rate of attrition among employees at a company will enable the management to respond promptly through the improvement of the internal policies and strategies of the company. In the situations where the skillful employees who have the potential to leave can be proposed a number of propositions, like a pay raise or adequate training, to minimize the chances of leaving. The employment of the machine learning models can assist companies to forecast the departure of employees. Analysts can use the historical data stored in the human resources (HR) departments to create and train a machine learning model capable of forecasting the departing employees of the company. These models are conditioned to analyze the relationship between the characteristics of both employees who are still working as well as those

who have already been terminated.

1.2 Objectives

The key aims of this study are as follows:

1.2.1 Develop a Machine Learning-Based System for Accurate Employee Attrition Prediction

The main aim of the project will be creating and executing an effective employee attrition prediction system that employs advanced machine learning algorithms like Support Vector Machine (SVM), Random Forest (RF) and K-Nearest Neighbors (KNN). The system will be fed with HR data that will provide a history to determine trends and relationships between employee characteristics and attrition behavior. The model is expected to provide high predictive accuracy and reliability by addressing the existing issues like the imbalance in the classes and irrelevant features, which are likely to be found in identifying employees at risk of leaving the company.

1.2.2 Enable Data-Driven Decision Support for Human Resource Management

In addition to prediction, the system can support the HR managers and organizational leaders in making proactive data-driven decisions. The organization can establish specific retention measures, including compensation increases, career advancement programs, workload balancing, or better workplace conditions by screening the high-risk employees beforehand. This goal will improve strategic workforce planning and minimize financial losses due to unplanned attritions.

1.2.3 Improve Model Performance Through Data Balancing and Feature Optimization

The other significant aim is to improve the performance of prediction by overcoming the real-world dataset issues that include: class imbalance and high dimensionality. Adaptive Synthetic Sampling (ADASYN) and feature selection methods will be used to enhance the minority class detection (attrition cases) and minimize the model complexity. This guarantees improved generalization, efficiency and predictability of the predictive system.

1.3 Scope

The scope of this research is defined by the following key areas:

1.3.1 Application of Supervised Machine Learning Techniques

The research involves application of the supervised learning algorithms such as SVM, Random Forest, and KNN to categorize the employees into attrition and non-attrition. The system measures various kernel functions and parameter configurations to find out the most efficient predictive model.

1.3.2 Handling Real-World HR Data Challenges

The typical problems of HR data that this study addresses are class imbalance, redundant features, and processing categorical data. Encoding, scaling, and normalization techniques are also used in order to have a better performance of the models. Improved prediction of attrition is achieved by using ADSYN and undersampling techniques.

1.3.3 Development of an Analytical Framework for HR Insights

The system is made not just to forecast the attrition but also to determine the most effective factors causing employees to leave their jobs, including overtime, monthly earnings, job grade and job satisfaction. This will help organizations to know about underlying causes and update internal policies to reflect the same.

1.3.4 Scalability and Future Expansion

The system will be such that it integrates with the real-time HR systems in future. It could be improved with the help of such advanced models as ANN and XGBoost and implemented as an HR analytics dashboard. The framework may also be generalized to other industries and feature sentiment analysis.

II. LITERATURE SURVEY

2.1 Traditional Methods For Studying Employee Attrition

Historically Previously researches were done by surveys and statistical analysis to establish factors such as job satisfaction, salary, and work environment impact on attrition. These were mostly descriptive and reactive. They assisted in observing reasons but were not very accurate when it comes to predicting future defections.

2.1.1 Reliance on Survey and Statistical Techniques

Conventional research relied on questionnaires and regression analysis to research turnover. These techniques listed correlations among variables but were not predictive. Organizations had the capability

of analyzing historical data but not predicting risk on an individual employee.

2.1.2 Sensitivity to Limited and Static Data

Majority of the initial researches employed small or organization-specific datasets. Missing or subjective data tended to influence results. This restricted their credibility and applicability to real-life.

2.1.3 Manual Feature Identification

Factors like pay, tenure and job satisfaction are picked manually by the researchers. It was time-consuming as it entailed domain knowledge. It was also prone to missing latent trends in big HR data.

2.1.4 Limited Scalability and Predictive Power

Conventional statistical models had a problem in the large and complex HR data. They were not able to manage the imbalance in classes. Consequently, they could not be used in the prediction of attrition on a large scale or in real-time.

2.2 Advances In Machine Learning For Employee Attrition Prediction

The current machine learning methods have enhanced attrition prediction through detection of intricate patterns in vast HR information. They are capable of treating more than one variable at a time unlike the traditional statistical models. This facilitates better and data-driven prediction of employee turnover.

Emergence of Ensemble and Advanced Models:Models such as Random Forest, XGBoost, and Artificial Neural Networks provide better prediction accuracy. These algorithms capture non-linear relationships between employee features. They also reduce overfitting and improve model reliability.

Handling Class Imbalance and Optimization Techniques: HR data tend to be smaller with the attrition cases than with non-attrition cases. The use of such techniques as ADASYN and SMOTE aids in the balancing of the data set. This enhances the identification of the employees who are facing threats of departure.

Comparative Performance Machine learning models are superior when compared to standard regression models. They are also more accurate and scalable. Such models are better applied in real-time HR analytics and decision-making.

2.3 Applications And Challenges Employee Attrition Prediction

Applications Across Domains: Employee attrition prediction is used in HR management and workforce planning to identify employees at risk of leaving. It helps organizations implement retention strategies and reduce recruitment costs. It also supports better strategic decision-making.

Challenges in Implementation: Attrition prediction faces challenges such as class imbalance, missing data, and changing employee behavior. Employee decisions are influenced by multiple complex factors. Data privacy and real-time integration also create implementation difficulties.

Need for Robust, Scalable Frameworks: Modern systems must handle large HR datasets efficiently and accurately. Proper preprocessing, feature selection, and data balancing are essential for reliable predictions. Scalable models integrated with real-time HR systems are important for practical use.

III. METHODOLOGY

3.1 Dataset Preparation

The dataset plays a crucial role in building an accurate and reliable employee attrition prediction system. This study uses the IBM HR Analytics dataset, which contains 1470 employee records with multiple features such as age, job role, monthly income, job satisfaction, overtime, total working years, and attrition status. The dataset includes both categorical and numerical attributes representing employee demographics, job-related factors, and organizational characteristics.

To prepare the data for model training, preprocessing steps were performed, including removal of irrelevant features, encoding categorical variables into numerical form, and applying feature scaling and normalization to ensure uniform data distribution. Since the dataset is imbalanced with fewer attrition (“Yes”) cases than non-attrition (“No”) cases, data balancing techniques such as ADASYN oversampling and undersampling were applied to improve minority class detection and enhance model performance.

3.2 System Architecture

The employee attrition prediction system will start with the employee dataset which will include demographic and job related attributes including; age, job role, salary, job satisfaction and overtime.

Exploratory Data Analysis (EDA) is conducted to attain an insight into trends and correlations, which is then followed by feature engineering tasks such as coding categorical data, scaling numerical data, and eliminating redundant features. As the dataset is skewed, resampling methods (SMOTE or ADASYN) are used to equalize between the cases of attrition and non-attrition. The resulting processed data is further divided into 80% training and 20% testing data and machine learning classifiers such as SVM, Random Forest, and KNN are trained avoiding hyperparameters to optimize performance. The most effective model is applied to predict the likelihood of an employee leaving or staying and the details are brought out on an HR analytics system to implement proactive retention initiatives.

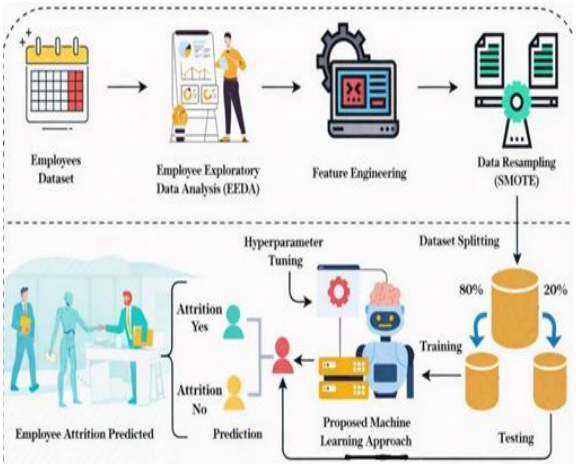

Figure 1: System Architecture

3.3 Machine Learning Model

The machine learning model serves as the core component of the employee attrition prediction system, responsible for classifying employees into "Attrition Yes" or "Attrition No" categories based on their features.

3.3.1 Model Architecture

In this study, machine learning algorithms such as Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN) were implemented for attrition classification. The input layer consists of processed employee features including age, monthly income, job satisfaction, overtime, job level, and total working years. After preprocessing and feature scaling, the data is fed into the selected classification model. Random Forest works by constructing multiple decision trees and combining their outputs to improve prediction accuracy, while SVM identifies the optimal decision boundary separating attrition and non-attrition cases. KNN classifies employees based on the majority class among their nearest neighbors. Feature selection techniques were also applied to reduce model complexity and improve generalization performance.

3.3.2 Compilation and Training

The dataset was split into training and testing sets (80:20 ratio), and the models were trained using the training data. Hyperparameter tuning was performed to optimize parameters such as the number of trees in Random Forest, kernel types in SVM, and the value of K in KNN. Model performance was evaluated using accuracy, precision, recall, and F1-score, with F1-score given special importance due to class imbalance. Cross-validation techniques were applied to ensure robustness and prevent overfitting, resulting in a reliable and efficient attrition prediction model.

3.4 Training And Validation

To test the performance of the model, the employee dataset was split in 80% training data and 20% testing (validation) data. Preprocessing procedures like scaling of features and class balancing were used during the training to facilitate robustness and fair learning between cases of attrition and non-attrition. The overfitting was avoided with cross-validation techniques, as well as to make the model applicable to the unseen data. Accuracy, precision, recall, and F1-score were used to evaluate the model performance, although special attention was paid to F1-score, because of the imbalance between the classes.

3.5 User Interface

The employee attrition prediction system has a user-friendly and interactive interface, which is to be used by HR professionals and non-technical users. The interface enables users to enter the information of the employees including age, job position, salary, job satisfaction, the presence of overtime and other applicable factors to come up with the prediction of attrition. It also offers clear options like Enter Employee Data and Predict Attrition and it even has simple instructions to follow. The system presents the prediction result in the form of Attrition Yes or No, and probability scores to make better decisions. The interface is friendly and compatible on both desktop

and web interfaces and it also has error handling features to check the input data and give significant responses in case of inappropriate or incomplete information is fed.

IV. IMPLEMENTATION

4.1 Tools And Technologies

To make the prediction system of employee attrition as efficient and accurate as possible, a mixture of data science tools and machine learning tools were employed to create the solution. The primary programming language was Python because its support of data analysis and machine learning libraries is large. The pre-processing and manipulation of the data were conducted with the help of Pandas and NumPy, whereas machine learning models (Random Forest, SVM, and KNN) were implemented with the help of Scikit-learn. Matplotlib and Seaborn were used to conduct data visualization and performance evaluation. ADASYN and SMOTE supporting libraries were used in order to deal with class imbalance. Flask was used to create a basic web based interface where users could key in the data of the employees and it would give them real time predictions. Such a combination of technology stack guaranteed a scalable and easy-to-use attrition prediction system.

4.2 Code Overview

The implementation of the Employee Attrition Prediction system is divided into three main parts:

4.2.1 Loading Data and Preprocessing

Pandas loads the dataset and analyses the dataset with regards to missing values or inconsistencies. Categorical variables are converted to numerical ones, irrelevant features are dropped and feature scaling is provided when needed. The dataset is then balanced with the help of ADASYN or SMOTE to correct the imbalance in the classes. Lastly, the data is divided into training and testing sets (80: 20).

4.2.2 Constructing and Training the Machine Learning Model

Scikit-learn is used to implement machine learning models including the Random Forest, SVM, and KNN. Hyperparameter optimization is used to optimize the model parameters and enhance prediction. The training dataset is fed to the models and is assessed with the help of accuracy, precision, recall, and F1-score. The use of cross-validation is done to stabilize the model and avoid overfitting.

4.2.3 Prediction and Classification

New employee information received by the web interface will go through the preprocessing phase (coding and scaling) before going through the trained model. This system then forecasts the chances of the employee to move in or out. The overall result of the prediction, as well as performance insights, is presented to help the HR managers make proactive retention decisions.

V. RESULT AND DISCUSSION

5.1 Model Performance

The employee attrition prediction model performed well during the testing with the best results obtained after balancing the dataset with the use of ADASYN. KNN model (K=3) with the maximum F1-score of approximately 0.93 performed the best whereas random forest and SVM also performed well. The confusion matrix showed that the majority of the employees were rightfully identified as either Attrition Yes or No, some few misclassifications were caused by the close nature of the employees. In general, data balancing and feature selection enhanced model accuracy and reliability to a great extent.

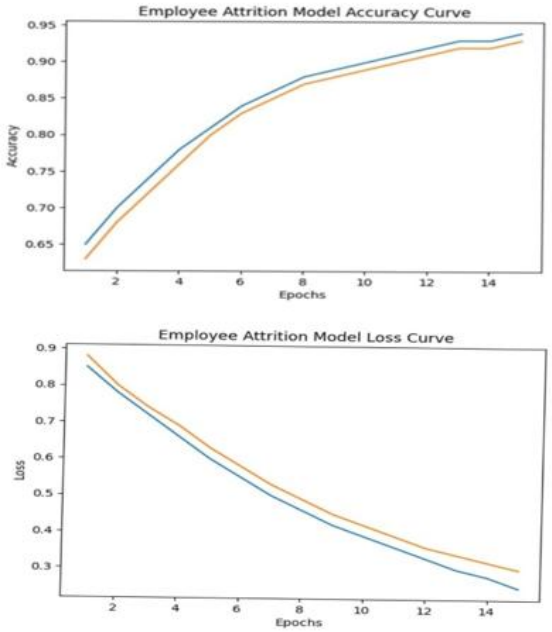

Figure 2 Training and Validation Accuracy

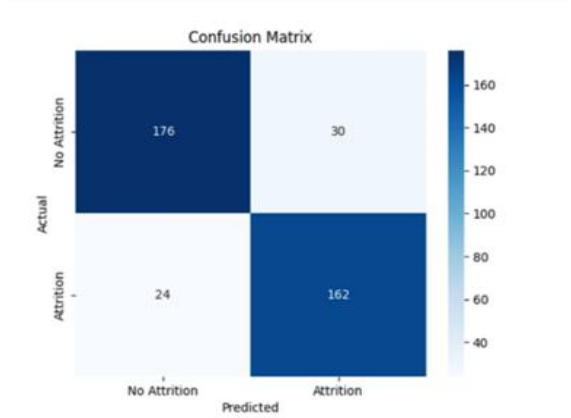

Figure Confusion Matrix

The confusion matrix of the prediction model of the employee attrition problem illustrated high precision, recall and F1-score of the prediction model of attrition Yes and attrition No which indicates high classification performance. The majority of the employees were rightfully classified particularly cases of non-attrition whereas few cases of misclassification were made where employees with similar characters (similar salaries, job designations or satisfaction rates) were incorrectly forecasted. This implies that despite the effectiveness of the machine learning models, there is still the possibility of the overlapping employee attributes creating some problems in prediction. To minimize such errors in future work, it is possible to introduce more behavioral features, real-time performance information, or more sophisticated ensemble or deep learning methods to enhance decision limits.

Figure 4 Output Screen

5.2 System Usability

The employee attrition prediction system showed good generalization ability during testing where the accuracy of the prediction system and F1-score became high following the application of data balancing methods like ADASYN. The training and validation performance measures were stable, which means that the model was not overfitting and was able to obtain some significant patterns within the HR data. The system was implemented in a easy and convenient web interface on which the HR professionals could key in the details of the employees and get immediate predictions. The lightweight machine learning models guaranteed fast processing time, and thus the system was applicable in real-time workforce analytics. Even though there have been some minor misclassifications because of the overlapping employee characters, the system was reliable and consistent in its performance, which is useful in decision making of the HR department.

5.3 Comparison With Traditional Methods

The conventional methods of analyzing employee attrition were based primarily on statistical methods, including regression analysis and survey-based research, which were aimed at determining the general factors affecting a turnover. Such techniques were manual in selection and did not have a great predictive power to single out individual at-risk employees. Conversely, the machine learning models used like the Random Forest, SVM and KNN automatically identify complex relationships in large HR datasets. They work better with high-dimensional data, class imbalance and nonlinear patterns. Consequently, machine learning-oriented solutions are more precise, scale better and are more proactive in delivering workforce management than the conventional statistical solutions.

5.4 Future Work

The further development of the employee attrition prediction system can be concerned with incorporating models more sophisticated like Artificial Neural Networks (ANN) and XGBoost to improve the accuracy of the prediction further. Real-time HR data, employee performance metrics and sentiment analysis with surveys or feedback systems could be incorporated to make a model more robust. The creation of an interactive HR analytics dashboard using visualization tools will increase in usability and decision support. Moreover, the extended system to cross-industry applications and the implementation of effective data privacy and security measures will make

the system more practical and reliable in its use in the organizational setting.

VI. CONCLUSION

Employee turnover is an important problem to the organizations because of the cost and effort involved in hiring skilled employees. This work reveals that machine learning models can results in a reasonable prediction of employee attrition in case of adequate data manipulation methods. ADASYN model balancing significantly increased model results, with KNN, Random forest, and SVM having high accuracy. The use of feature selection also made the model less complex with a high predictive accuracy, but manual undersampling lowered the performance because vital information was lost. Moreover, the proposed system can be easily scaled to large sets of employees, as well as make decisions quicker and using data. On the whole, this method is an effective and credible solution to assist the organization to recognize the possibility of attrition at the first stages and to take the appropriate measures that would ensure employee retention.

REFERENCES

[1] S. Kaur and R. Vijay, "Job satisfaction – A major factor behind attrition or retention in the retail industry," Imperial Journal of Interdisciplinary Research, vol. 2, no. 8, 2016.

[2] D. G. Gardner, L. V. Dyne, and J. L. Pierce, "The effects of pay level on organization-based self-esteem and performance: A field study," Journal of Occupational and Organizational Psychology, vol. 77, no. 3, pp. 307–322, 2004.

[3] E. Moncarz, J. Zhao, and C. Kay, "An exploratory study of U.S. lodging properties' organizational practices on employee turnover and retention," International Journal of Contemporary Hospitality Management, vol. 21, no. 4, pp. 437–458, 2009.

[4] Q. A. Al-Radaideh and E. A. Nagi, "Using data mining techniques to build a classification model for predicting employees' performance," International Journal of Advanced Computer Science and Applications, vol. 3, no. 2, pp. 144–151, 2012.

[5] G. K. P. V. Vijaya Saradhi and B. S. Palshikar, "Employee churn prediction," Expert Systems with Applications, vol. 38, no. 3, pp. 1999–2006, 2011.

[6] S. Rogers and M. Girolami, A First Course in Machine Learning. Boca Raton, FL, USA: CRC Press, 2016.

[7] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[8] N. Cristianini and B. Schölkopf, "Support vector machines and kernel methods: The new generation of learning machines," AI Magazine, vol. 23, no. 3, p. 31, 2002.

[9] T. K. Ho, "Random decision forests," in Proc. 3rd Int. Conf. Document Analysis and Recognition (ICDAR), 1995.

[10] X. Zhu, Knowledge Discovery and Data Mining: Challenges and Realities. Hershey, PA, USA: IGI Global, 2007.

[11] D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," Journal of Machine Learning Technologies, 2011.

[12] H. He and E. A. Garcia, "Learning from imbalanced data," IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263–1284, 2009.

[13] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in Proc. IEEE Int. Joint Conf. Neural Networks (IJCNN), 2008.

[14] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," Journal of Machine Learning Research, vol. 3, pp. 1157–1182, 2003.

[15] W. Zhu, X. Wang, Y. Ma, M. Rao, J. Glimm, and J. S. Kovach, "Detection of cancer-specific markers amid massive mass spectral data," Proceedings of the National Academy of Sciences, vol. 100, no. 25, pp. 14666–14671, 2003.