A Hybrid Case-Based Reasoning Framework Using KNN, Word2Vec, and Cosine Similarity for Employee Attrition Analysis

Akhmad Arif Faisal Siregar[1], Ema Utami[2*] , Tika Novita Sari[3]

[1,2] *Department of Informatics, Universitas Amikom Yogyakarta, Yogyakarta, Indonesia*

*corr-author: ema.u@amikom.ac.id

Abstract - Employee attrition prediction remains a longstanding challenge in human resource analytics, as organizations increasingly depend on computational decision-support systems that are transparent, consistent, and operationally accountable. Conventional methods that rely solely on numerical attributes are restricted in their ability to accurately capture the structural and contextual relationships inherent in categorical and text-based employee descriptors. To overcome this limitation, the current study investigates a hybrid Case-Based Reasoning (CBR) retrieval framework that combines K-Nearest Neighbors (KNN) with Word2Vec embeddings derived from the dataset's limited textual attributes, specifically Department, Gender, EducationField, MaritalStatus, and OverTime. Eight experimental configurations were assessed to examine the impact of alternative similarity metrics and diverse feature representations. The optimal configuration of KNN, enhanced with Word2Vec embeddings and cosine similarity, attained an accuracy of 0.8526 and a weighted F1-score of 0.8000, thereby exceeding the performance of baseline models based solely on numerical features and those utilizing Manhattan distance. Nonetheless, the improvements in performance remained limited owing to dataset-specific limitations, such as class imbalance and the inherently superficial characteristics of the textual descriptors, which restrict the semantic richness of Word2Vec embeddings. Furthermore, the IBM attrition dataset does not encompass downsizing or termination situations, highlighting conceptual and ethical constraints when utilizing similarity-based predictions for high-stakes HR decisions. Overall, the findings indicate that hybrid similarity representations, particularly the combination of Word2Vec embeddings with cosine distance, can improve the structural expressiveness of CBR, although their predictive effectiveness is still limited by data sparsity and considerations of fairness.

Keywords: Case-Based Reasoning; employee attrition; K-Nearest Neighbors; Word2Vec; human resource analytics.

I. INTRODUCTION

Human resource management has evolved into an increasingly data-driven function as organizations aim to forecast employee behavior and enhance workforce planning [1]. n Indonesia, with over 143 million individuals engaged in the labor force and micro, small, and medium enterprises (MSMEs) representing the predominant sector of business entities, workforce stability is essential for organizational resilience and competitiveness [1]. Global economic dynamics, policy uncertainty, and swift technological advancements pose considerable challenges for Indonesian companies, especially regarding operational efficiency and competitiveness [1].

Within this framework, human resource development (HRD) has emerged as a critical element in securing the long-term sustainability of the organization. Predictive analytics in human resource management is progressively acknowledged as a vital instrument in talent management and strategies for employee retention [2]. Considering the substantial expenses associated with recruitment and their influence on morale and productivity, the implementation of effective retention strategies is essential. Previous research emphasizes that utilizing data-driven analytics not only enhances recruitment efficiency but also reinforces employee loyalty [3].

Conventional statistical models, including logistic regression and decision trees, have been extensively employed to analyze attrition; however, they frequently encounter difficulties in effectively modeling nonlinear and high-dimensional relationships within employee data [4]. With the progress of machine learning, more advanced models such as Random Forest, Gradient Boosting, XGBoost, and Artificial Neural Networks have exhibited superior predictive accuracy, with reported success rates often surpassing 90% in benchmark evaluations[5]. With the progression of machine learning techniques, more advanced models

such as Random Forest, Gradient Boosting, XGBoost, and Artificial Neural Networks have exhibited enhanced predictive accuracy, often surpassing 90% in benchmark evaluations [6]. Recent studies have also investigated explainable AI (XAI) methodologies to enhance transparency in HR analytics, such as SHAP-based feature attribution and interpretable decision rules. These advancements suggest a transition toward models that not only demonstrate high performance but also provide transparent justifications for their predictions, thereby supporting organizational accountability [7].

Even with these improvements, there are still some problems. First, a lot of the research mostly uses numbers and categories, so it doesn't look at the possible worth of text-based information like job roles, department descriptions, or qualitative reviews. Deep learning models can capture semantic linkages, but it is not yet clear how well they interact with similarity-based reasoning frameworks like Case-Based Reasoning (CBR) [8, 9]. Second, the Kaggle IBM HR Attrition dataset, which is extensively used, has only a little bit of textual depth. This makes it hard to figure out how meaningful semantic representations may be made from small categorical variables. Third, even though attrition prediction is often seen as a way to improve the whole workforce, the dataset itself doesn't show situations where people are laid off, downsized, or fired based on their performance. These conceptual deficiencies underscore the necessity for meticulous methodological positioning and ethical delineation in the application of predictive models to sensitive human resource choices [10].

To tackle these issues, this research examines a hybrid CBR framework that combines K-Nearest Neighbors (KNN) with Word2Vec embeddings, one-hot encoded categorical features, and cosine similarity. CBR provides a comprehensible framework for assessing new cases by comparing them to prior employee profiles, facilitating transparency in retrieval-based decision support. The paper investigates the potential of lightweight semantic representations to improve similarity computation through the integration of categorical fields via Word2Vec, especially when used alongside numerical characteristics and various distance metrics.

This study offers three contributions. First, it systematically compares eight hybrid similarity models, from KNN, which is simply numerical, to models that combine OHE, Word2Vec, Manhattan distance, and cosine similarity. Second, it offers empirical evidence about the degree to which Word2Vec enhances retrieval accuracy within a dataset comprising solely superficial textual elements [11]. Third, it critically assesses the practical and ethical constraints of employing CBR for HR analytics, highlighting the differences between attrition modelling and decision-making scenarios, such downsizing or employee selection.

The primary objective of this work is to enhance comprehension of similarity-based reasoning in HR analytics by evaluating the advantages and limitations of incorporating semantic and numerical representations inside KNN-driven CBR retrieval. The findings provide insights into the practicality, interpretability, and hazards linked to the implementation of such models in actual organizational contexts.

A. Related Works

Case-Based Reasoning (CBR) has been widely implemented as a decision-support framework across a variety of disciplines [8] [9]. Within the legal sphere, Lopes et al. (2009) utilized Case-Based Reasoning (CBR) to develop a sentencing recommendation system in Brazil[12]. Within the domain of logistics, CBR has been employed to facilitate emergency response initiatives and to underpin intelligent decision-support systems [13]. In the domain of artificial intelligence research, Martins and Neto (2020) combined Case-Based Reasoning with Self-Organizing Maps to tackle intricate decision-making tasks [14]. Within the financial sector, CBR-based knowledge systems have been employed to facilitate commercial loan recommendations [15], whereas in the domain of aviation maintenance, Naqvi et al. (2022) integrated case-based reasoning with Natural Language Processing through BERT embeddings to enhance case matching capabilities [16]. In recent developments, Case-Based Reasoning has been utilized within the domain of human resource management to forecast project workload and personnel requirements [17].

Research on employee retention has expanded significantly in recent years, employing diverse approaches to elucidate the factors that influence employees' decisions to remain with an organization. Several machine learning-based methodologies have been employed, including Logistic Regression (LR) [18], Decision Tree (DT) [6], Random Forest (RF) [18], Support Vector Machine (SVM) [18], K-Nearest Neighbors (KNN) [6, 19], Ensemble Learning [6], and Gradient Boosting [10]. Furthermore, deep learning models, including Deep Neural Networks (DNN)[20], Convolutional Neural Networks (CNN) [20], and Multi-Layered Neural Networks [5], have also been employed to enhance the precision of employee retention predictions.

Bongale et al. (2023) utilized Logistic Regression and achieved an accuracy of 72%, although the performance was limited by the model's incapacity to identify nonlinear patterns[21]. Sharma and Singla (2024) employed multi-layer neural networks, attaining 91% accuracy and surpassing LR and DT models, though at a higher computational expense [5]. Muthugala et al. (2024) evaluated RF, SVM, and DT, identifying Random Forest as the most effective, achieving a retention accuracy of 94.5% [18]. Silpa et al. (2023) demonstrated that the integration of advanced feature selection with XGBoost achieved an accuracy of up to 95% [22]. Pandey et al. (2024) identified Random Forest as the most effective method among LR, RF, and Gradient Boosting, attaining an accuracy of 85% [10]. Mitravinda et al. (2022) combined KNN with collaborative filtering and identified XGBoost as the superior model, achieving an accuracy of 87.07% [23]. Yashu et al. (2024) enhanced performance further by employing DNN and CNN models, attaining an accuracy of up to 95% [20].

Recent research on employee attrition prediction has utilized a range of machine learning and deep learning algorithms. Traditional classifiers such as Logistic Regression, Random Forest, and XGBoost have exhibited excellent predictive capabilities owing to their robustness when applied to structured HR data. Recent methods employ deep neural networks and ensemble learning techniques to effectively model the nonlinear relationships among employee attributes. Concurrently, explainable AI (XAI) methodologies have been developed to enhance transparency within HR decision support systems.

Nonetheless, the majority of current research emphasizes prediction accuracy over case-based interpretability or reasoning prompted by similarity. Few studies investigate the integration of Case-Based Reasoning (CBR) with semantic representation to facilitate analogical decision-making within human resources environments. This research addresses this lacuna by combining text-based semantic similarity with case-based reasoning, facilitating interpretable and experience-oriented recommendations beyond simple classification.

II. METHOD

This study adopts an experimental framework designed to evaluate the effectiveness of a hybrid Case-Based Reasoning (CBR) retrieval model that integrates numerical attributes and Word2Vec-derived semantic embeddings within a cosine-based similarity space. The methodological components are outlined in the subsections below.

A. Dataset

The study employs the publicly available IBM HR Attrition dataset consisting of 1,470 employee records and 34 attributes. The target variable Attrition contains two classes ("Yes" and "No"), with a strong imbalance favoring the "No" class. The dataset includes demographic variables, work environment indicators, compensation-related fields, and employment history details. The features are grouped into three categories:

1) Numerical Attributes: Age, DailyRate, DistanceFromHome, Education, EmployeeCount, EmployeeNumber, EnvironmentSatisfaction, HourlyRate, JobInvolvement, JobLevel, JobSatisfaction, MonthlyIncome, MonthlyRate, NumCompaniesWorked, PercentSalaryHike, PerformanceRating, RelationshipSatisfaction, StandardHours, StockOptionLevel, TotalWorkingYears, TrainingTimesLastYear, WorkLifeBalance, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager.

2) Text-Derived Attributes: Department, Gender, EducationField, MaritalStatus, and OverTime (which are treated as textual inputs for embedding).

*3) Other categorical features***:** BusinessTravel, JobRole, Over18.

B. Data Preprocessing

Missing numerical values are imputed using median substitution, while missing textual entries are replaced with an empty token prior to embedding. Numerical attributes are standardized using z-score normalization to ensure consistent feature scaling in similarity computation. The dataset is partitioned into training and testing sets with a stratified 70/30 split to preserve class distribution.

C. Categorical Encoding via One-Hot Encoding

Categorical attributes (*BusinessTravel, JobRole, Over18*) are transformed using One-Hot Encoding (OHE). OHE preserves semantic neutrality by treating categories as orthogonal dimensions, enabling the model to measure similarity based on exact category matches. The OHE matrix is integrated into the final feature representation alongside numerical and embedded textual vectors.

D. Oversampling with SMOTE

To mitigate the severe imbalance in the *Attrition = Yes* class, the Synthetic Minority Over-sampling Technique (SMOTE) is applied selectively in extended experiments. SMOTE generates synthetic samples in the

minority class using k-nearest neighbor interpolation, improving classifier recall. SMOTE is applied after full feature transformation, ensuring synthetic samples represent the combined numerical, categorical, and embedding spaces consistently.

E. Distance Metrics for Similarity Computation

Because CBR relies on retrieving the most similar historical cases, the choice of distance metric is central. Two families of metrics are evaluated:

1) Manhattan Distance: used for baseline numerical and hybrid models. It computes absolute differences across features and is robust to outliers as in (1).

$$d_{\text{Manhattan}}(\text{X}, \text{Y}) = \sum_{i=1}^{n} |x_i - y_i| \qquad (1)$$

where x and y denote two text-based feature vectors representing employee cases, and the similarity between them is computed using cosine-based measurement.

2) Cosine Distance: applied to the hybrid Word2Vec–numeric representations. Cosine distance measures the angular difference between high-dimensional vectors, making it suitable for embeddings. The custom cosine function is defined as:

$$\cos(X, Y) = 1 - \frac{x \cdot y}{\|x\| \|y\|} \qquad (2)$$

where x and y denote two text-based feature vectors representing employee cases, and the similarity between them is computed using cosine-based measurement. This metric is compatible with heterogeneous feature spaces combining dense embeddings and numerical scaling.

F. Evaluation Metrics

Model performance is assessed using accuracy, precision, recall, F1-score, and confusion matrix analysis to provide a comprehensive evaluation of classification effectiveness and error distribution.

G. Proposed Model

The suggested system's foundation is based on Case-Based Reasoning (CBR), which organizes the decision-support process into four standard stages: Retrieve, Reuse, Revise, and Retain. Fig. 1 illustrates the comprehensive workflow of the proposed hybrid Case-Based Reasoning (CBR) retrieval system, which amalgamates numerical characteristics with Word2Vec-derived textual embeddings within a cohesive cosine similarity space.

During the Retrieve stage, each new case is evaluated against previously stored cases utilizing a hybrid text-based similarity approach. Limited textual attributes within the dataset, specifically Department, EducationField, Gender, MaritalStatus, and OverTime are initially concatenated into a single textual representation for each case. The text is subsequently normalized via tokenization and fundamental cleaning, then transformed into semantic vectors using Word2Vec trained from start with the Skip-gram architecture (vector size = 200, window = 5, epochs = 15). A document-level embedding is derived through mean pooling of token vectors.

Case similarity is determined through cosine distance within the embedding space, which is appropriate for high-dimensional representations. The K-Nearest Neighbors (KNN) algorithm (k = 31, distance-weighted voting, brute-force search) is utilized to identify the most comparable historical cases and to determine the attrition designation for the new case.

The Reuse stage utilizes the predicted outcome from the retrieved neighbors, whereas the Revise stage permits optional modifications when similarity is inadequate or expert intervention is necessary. Finally, the Retain stage facilitates the storage of new cases and finalized decisions to enhance the case base for subsequent retrieval.

III. RESULT AND DISCUSSION

A. Performance Comparison

Table II summarizes the weighted accuracy, precision, recall, and F1-score for all eight KNN-based CBR retrieval configurations.

The findings reveal minor yet constant performance variations among encoding techniques and distance measurements. Models utilizing cosine distance typically surpass those based on Manhattan distance, although configurations that integrate Word2Vec embeddings attain somewhat superior F1-scores compared to label-encoded representations. The incorporation of One-Hot Encoding (OHE) does not produce a significant enhancement in performance relative to models lacking OHE, since both configurations exhibit similar retrieval efficacy. Considering that One-Hot Encoding significantly elevates feature dimensionality and computing expenses, the proposed method intentionally omits OHE, employing a KNN + Word2Vec + Cosine framework to preserve comparable predictive performance while enhancing computational efficiency and model simplicity.

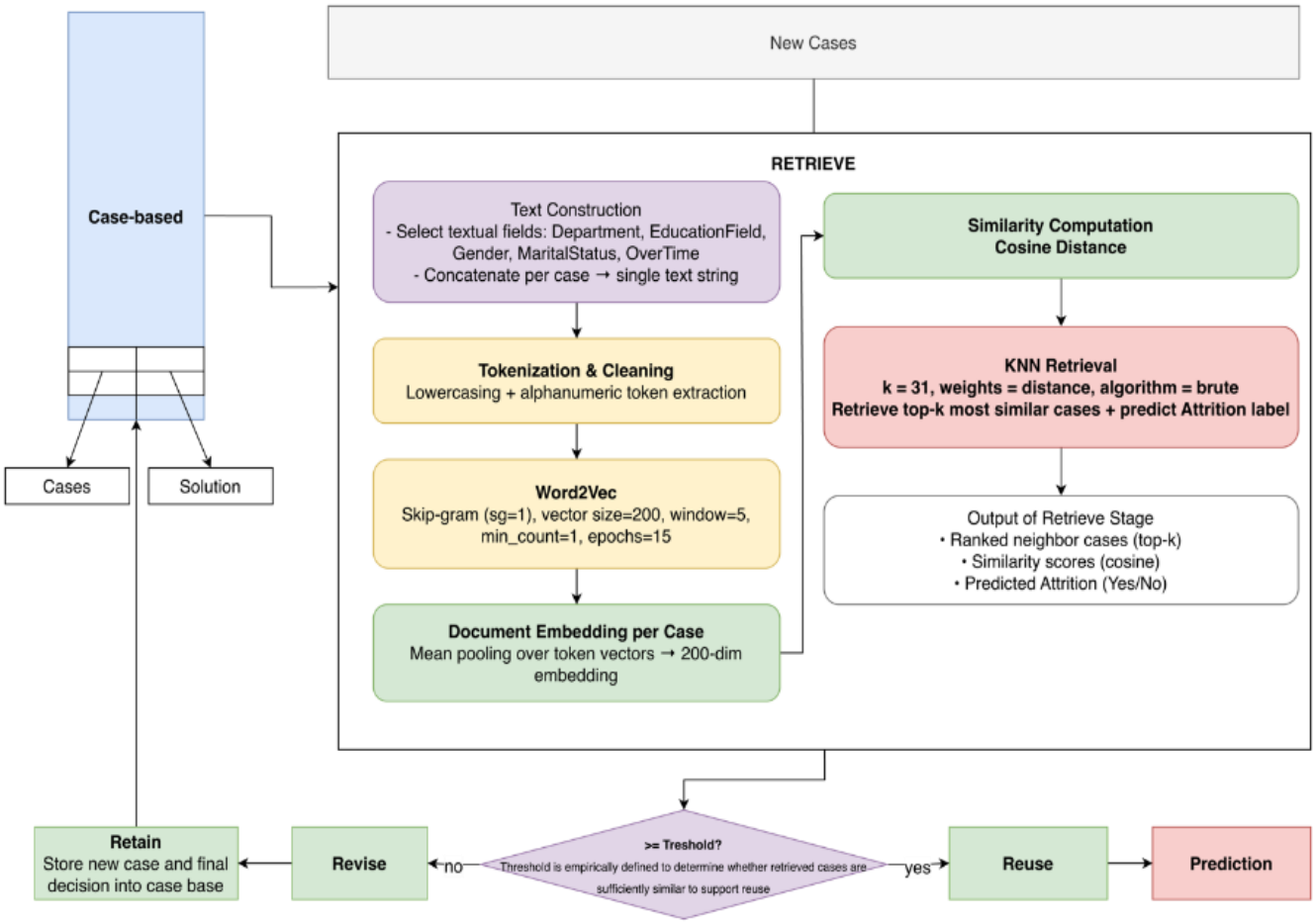

Fig. 1 Hybrid case-based rasoning retrieval framework

B. The Impact SMOTE for Confusion Matrix Analysis

To gain a more comprehensive comprehension of classification performance, confusion matrices were produced for the proposed model and its hybrid variants, both prior to and following the application of SMOTE oversampling. The analysis concentrates on the minority class (Attrition = Yes), which represents approximately 14% of the dataset.

1) Confusion Matrix of the Proposed Model without Oversampling (Fig. 2)

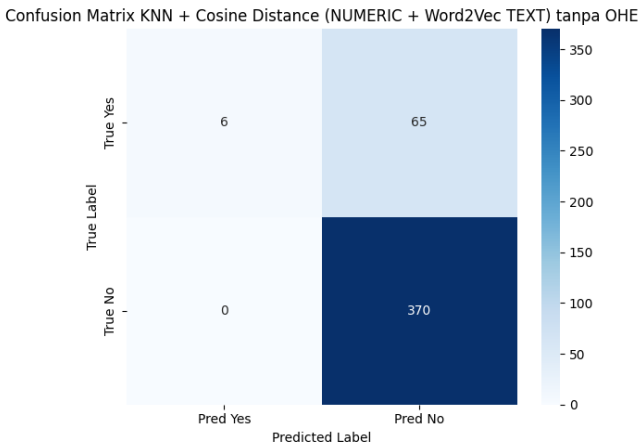

Fig. 2 Confusion matrix of the proposed model without oversampling

TABLE II
MODEL PERFORMANCE COMPARISON

Model	Accuracy	Precision	Recall	F1-score
KNN + Manhattan	0.8458	0.8700	0.8500	0.7800
KNN + Manhattan + Word2Vec	0.8458	0.8400	0.8500	0.7700
KNN + Manhattan + Word2Vec + OHE	0.8458	0.8700	0.8500	0.7800
KNN + Cosine	0.8503	0.8700	0.8500	0.7900
KNN + Cosine + Word2Vec (Proposed)	**0.8526**	**0.8700**	**0.8500**	**0.8000**
KNN + Cosine + Word2Vec + OHE	0.8526	0.8700	0.8500	0.8000
KNN + Cosine + Word2Vec + SMOTE	0.5600	0.7700	0.5600	0.6200
KNN + Cosine + Word2Vec + OHE + SMOTE	0.5800	0.7800	0.5800	0.6300

The model exhibits an exceptionally high true negative rate, accurately identifying all "No" cases, but incorrectly classifies the majority of minority instances, with only six correctly predicted. This illustrates the impact of both class imbalance and KNN's density-based neighborhood framework.

2) Confusion Matrices of the Proposed Model with SMOTE Oversampling. To address severe imbalance, SMOTE was employed on the training set prior to model development.

The implementation of SMOTE substantially enhances the model's capacity to recognize minority-class instances, as evidenced by an increase in true positive predictions from 6 to 41. Concurrently, this enhancement is associated with a considerable increase in false positives, from zero to 163, thereby signifying a notable decline in specificity. The results indicate that although SMOTE enhances KNN's ability to identify minority-class patterns, the synthetic interpolation within high-dimensional feature spaces may unduly distort class boundaries, thereby promoting overgeneralization and increasing the misclassification of majority-class instances.

Consistent trends are observed across all experimental combinations. The original non-SMOTE model has high specificity but is ineffective in identifying minority-class instances, highlighting a recognized shortcoming of KNN in cases of significant class imbalance. While SMOTE enhances recollection by augmenting minority-class identification, it concurrently leads to a significant reduction in precision, creating an undesirable trade-off for decision-support systems. Moreover, the textual features encoded by Word2Vec provide minimal discriminative value due to the limited word count of the sample and the lack of really unstructured textual tales. These findings collectively suggest that the principal performance limitation stems from dataset constraints rather than the retrieval mechanism, underscoring the necessity for more comprehensive HR textual data or improved feature representations in future research to enhance semantic separability and predictive fairness/

The comparative assessment of the proposed CBR–KNN retrieval models indicates that while performance variations among the Manhattan, Cosine, OHE, and Word2Vec variants are observed, they remain consistently modest. This pattern indicates inherent structural constraints of both the dataset and the similarity-based retrieval methods, rather than suboptimal parameter configurations. Analysis of the confusion matrix indicates that all configurations predominantly predict the majority class, while facing challenges in accurately identifying minority (Attrition = Yes) cases, a behavior theoretically anticipated under conditions of significant class imbalance in distance-based classifiers.

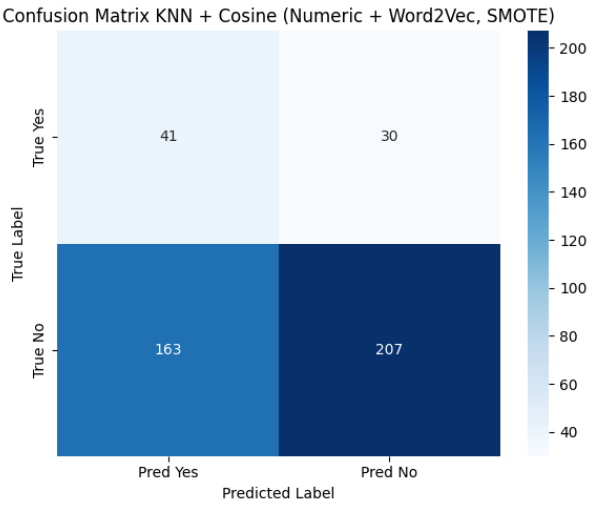

Fig. 3 Confusion matric of the proposed model with oversampling

Cosine distance exhibits marginally greater robustness than Manhattan distance in hybrid representations, especially when dealing with high-dimensional embeddings. Nevertheless, these improvements remain limited, suggesting that the choice of distance metric alone is insufficient to address the issue of minority-class sparsity. Word2Vec offers only marginal enhancement due to the textual attributes comprising concise categorical identifiers (e.g., department, marital status, overtime) rather than semantically comprehensive narratives. Consequently, embeddings function predominantly as stabilizing mechanisms rather than as providers of novel discriminative information.

Compared to leading models such as Random Forest, XGBoost, and deep neural networks documented in previous research, the proposed approach does not seek to attain higher predictive accuracy. Rather, its contribution resides in transparent, case-based retrieval that facilitates human-centered decision analysis, emphasizing the trade-off between interpretability and predictive accuracy. Oversampling experiments employing SMOTE affirm this limitation: while recall experiences significant enhancement, the incidence of false positives rises markedly, thereby raising ethical and operational considerations in HR decision-making scenarios.

From an ethical perspective, the utilization of CBR in workforce analytics involves potential risks of perpetuating biases and compromising fairness, as similarity retrieval may reinforce historical organizational patterns incorporated in prior cases. Therefore, the proposed model is intended solely as a decision-support instrument that necessitates human supervision, rather than functioning as an automated system for downsizing or termination determinations.

IV. CONCLUSION

This research assessed a hybrid Case-Based Reasoning framework that incorporates numerical attributes, one-hot encoded categorical features, and Word2Vec embeddings within a K-Nearest Neighbors retrieval procedure. Eight model configurations were evaluated utilizing Manhattan and cosine distance metrics, both with and without the incorporation of semantic embeddings. The findings suggest that, although the hybrid models offer marginal enhancements compared to the exclusively numerical baselines, the improvements are limited across all assessment criteria. This observation implies that the predictive accuracy of similarity-based retrieval is limited not solely by algorithmic architecture, but also by the representational boundaries inherent within the dataset. A primary observation is that the Kaggle attrition dataset, while extensively utilized, exhibits restricted textual richness, thereby causing Word2Vec to function more as an alternative encoding technique than a true semantic model. As a result, the hybrid methodology improves numerical continuity, yet it is unable to substantially alter the decision space. Moreover, the pronounced class imbalance systematically results in under-detection of the minority class within non-oversampled models. Conversely, oversampling techniques enhance recall, albeit at the cost of a significant increase in false positives. These trade-offs underscore the pragmatic constraints inherent in employing k-nearest neighbor-based case-based reasoning for operational decision-making within the domain of human resources analytics. This study recognizes certain conceptual limitations. Specifically, attrition labels do not invariably align with retrenchment decisions, and the dataset reflects historical trends that may incorporate demographic or structural biases. Consequently, the implementation of retrieval-based recommendations within actual organizational contexts necessitates thorough fairness assessments, transparent model interpretation, and human oversight to mitigate the potential for perpetuating inequitable results. In conclusion, the presented hybrid case-based reasoning model indicates that the integration of distributed embeddings within similarity-based reasoning can produce incremental enhancements; however, its utility is constrained by data quality, class imbalance, and ethical considerations. Future research should utilize datasets incorporating more comprehensive textual data, investigate adaptive similarity functions or metric learning methodologies, and incorporate fairness-aware mechanisms to enhance predictive accuracy and facilitate responsible implementation within HR decision support systems.

REFERENCES

[1] Badan Pusat Statistik, “Statistik Indonesia 2023,” 2023. [Online]. Available: https://www.bps.go.id/id/publication/2023/02/28/18018f9896f09f03580a614b/statistik-indonesia-2023.html

[2] N. Yahia, J. Hlel, and R. Colomo-Palacios, “From Big Data to Deep Data to Support People Analytics for Employee Attrition Prediction,” *IEEE Access*, vol. 9, pp. 60447–60458, 2021, doi: 10.1109/ACCESS.2021.3082391.

[3] S. Paigude, S. C. Pangarkar, S. N. Hundekari, M. Mali, K. Wanjale, and Y. Dongre, “Potential of Artificial Intelligence in Boosting Employee Retention in the Human Resource Industry,” *International Journal on*

Recent and Innovation Trends in Computing and Communication, 2023.

[4] T. S. I and T. Saranya, "Forecast Of Employee Attrition In Big Data To Support People Analytics," *International Journal of Scientific Research In Engineering And Management*, 2023.

[5] R. Sharma and A. Singla, "Deep Learning in HRM: Transforming Employee Retention through Predictive Analytics," in *2024 4th Asian Conference on Innovation in Technology (ASIANCON)*, 2024, pp. 1–6. doi: 10.1109/ASIANCON62057.2024.10837776.

[6] B. Kaur and A. Dogra, "A Machine Learning Model for Predicting Employees Retention: An Initiative towards HR through Machine," in *2022 Seventh International Conference on Parallel, Distributed and Grid Computing (PDGC)*, 2022, pp. 653–657. doi: 10.1109/PDGC56933.2022.10053249.

[7] G. Marín Díaz, J. J. Galán Hernández, and J. L. Galdón Salvador, "Analyzing Employee Attrition Using Explainable AI for Strategic HR Decision-Making," *Mathematics*, vol. 11, no. 22, 2023, doi: 10.3390/math11224677.

[8] A. Aamodt and E. Plaza, "Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches," AI Communications. IOS Press, 1994.

[9] I. D. Watson and F. Marir, "Case-based reasoning: A review," *Knowl Eng Rev*, vol. 9, pp. 327–354, 1994, [Online]. Available: https://api.semanticscholar.org/CorpusID:41059740

[10] D. K. Pandey, S. Upadhyay, A. K. Jha, S. Rana, and M. Singh, "Leveraging HR Analytics for Predictive Talent Management and Employee Retention," in *2024 13th International Conference on System Modeling & Advancement in Research Trends (SMART)*, 2024, pp. 436–440. doi: 10.1109/SMART63812.2024.10882581.

[11] D. Srivamsi, O. M. Deepak, M. D. A. Praveena, and A. Christy, "Cosine Similarity Based Word2Vec Model for Biomedical Data Analysis," in *7th International Conference on Trends in Electronics and Informatics, ICOEI 2023 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 1400–1404. doi: 10.1109/ICOEI56765.2023.10125794.

[12] E. C. Lopes, U. Schiel, and G. P. dos Santos Jr., "A Decision Support Methodology for the Control of Alternative Penalties - A Case-Based Reasoning Approach," in *2009 International Conference on Information, Process, and Knowledge Management*, 2009, pp. 72–77. doi: 10.1109/eKNOW.2009.19.

[13] X.-M. Han and J.-T. Han, "A Research of Intelligent Decision Support System of ELS^3 Based on Case-Based Reasoning," in *2007 International Conference on Wireless Communications, Networking and Mobile Computing*, 2007, pp. 5765–5768. doi: 10.1109/WICOM.2007.1413.

[14] D. M. L. Martins and F. B. de Lima Neto, "Hybrid Intelligent Decision Support Using a Semiotic Case-Based Reasoning and Self-Organizing Maps," *IEEE Trans Syst Man Cybern Syst*, vol. 50, no. 3, pp. 863–870, 2020, doi: 10.1109/TSMC.2017.2749281.

[15] A. Adla and M. Frendi, "A Decision Support Systemfor Commercial Lending," in *2021 International Conference on Decision Aid Sciences and Application (DASA)*, 2021, pp. 326–331. doi: 10.1109/DASA53625.2021.9682296.

[16] S. M. R. Naqvi, M. Ghufran, S. Meraghni, C. Varnier, J.-M. Nicod, and N. Zerhouni, "CBR-Based Decision Support System for Maintenance Text Using NLP for an Aviation Case Study," in *2022 Prognostics and Health Management Conference (PHM-2022 London)*, 2022, pp. 344–349. doi: 10.1109/PHM2022-London52454.2022.00067.

[17] O. Kovalchuk, D. Kobylkin, and O. Zachko, "HR Decision-Making Support System Based On The CBR Method," in *2023 IEEE 18th International Conference on Computer Science and Information Technologies (CSIT)*, 2023, pp. 1–4. doi: 10.1109/CSIT61576.2023.10324169.

[18] D. Muthugala, S. M. Arachchi, P. Pallewatta, A. Maithripala, and G. Seneviratne, "Predicting Employee Attrition & Employee Retention Period using Supervised Learning," in *2024 6th International Conference on Advancements in Computing (ICAC)*, 2024, pp. 127–132. doi: 10.1109/ICAC64487.2024.10851009.

[19] M. Muhammad, T. Sutikno, and I. Riadi, "A Comparative Study of K-Means and KNN Imputation for Handling Missing Data in Scholarship Applicant Datasets," *JUITA: Jurnal Informatika*, vol. 13, no. 3, pp. 245–254, Nov. 2025, doi: 10.30595/juita.v13i3.26502.

[20] Yashu, R. Sharma, A. Jain, and M. Manwal, "Enhancing Human Resource Management through Deep Learning: A Predictive Analytics Approach to Employee Retention Success," in *2024 IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems (ICITEICS)*, 2024, pp. 1–4. doi: 10.1109/ICITEICS61368.2024.10625175.

[21] A. M. Bongale, D. Dharrao, and S. Urolagin, "Exploratory Data Analysis and Classification of Employee Retention based on Logistic Regression Model," in *2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS)*, 2023, pp. 1929–1933. doi: 10.1109/ICACCS57279.2023.10112681.

[22] N. Silpa, V. V. R. Maheswara Rao, M. V. Subbarao, R. R. Kurada, S. S. Reddy, and P. J. Uppalapati, "An Enriched Employee Retention Analysis System with a Combination Strategy of Feature Selection and Machine Learning Techniques," in *2023 7th International Conference on Intelligent Computing and Control*

Systems (ICICCS), 2023, pp. 142–149. doi: 10.1109/ICICCS56967.2023.10142473.

[23] K. M. Mitravinda and S. Shetty, "Employee Attrition: Prediction, Analysis Of Contributory Factors And Recommendations For Employee Retention," in *2022 IEEE International Conference for Women in Innovation, Technology & Entrepreneurship (ICWITE)*, 2022, pp. 1–6. doi: 10.1109/ICWITE57052.2022.10176235.