Interpretable Temporal Risk Modeling for Contributor Inactivity Prediction: A Comparative Study of Tree-Based Ensembles

Adi Suryaputra Paramita[1,*], Indra Maryati[2], Christian[3], Elizabeth Nathania Witanto[4], Auezova Raya Tileubaevna[5], Choo Wou Onn[6]

[1,2,3,4]*School of Information Technology, Universitas Ciputra, CBD Boulevar, Citraland Surabaya 60219, Indonesia*

[5]*Department of Software Engineering of the University of Innovation Technologies, Uzbekistan*

[6]*Faculty of Data Science and Information Technology, INTI International University, Malaysia*

(Received: October 26, 2025; Revised: December 10, 2025; Accepted: March 1, 2026; Available online: April 26, 2026)

Abstract

This study aims to develop an interpretable temporal risk modeling framework for predicting contributor inactivity in collaborative development environments, thereby supporting sustained participation and improving productivity. The research focuses on contributor activity data collected from a collaborative software development platform, in which participation histories are represented by temporal engagement features that capture activity recency, participation intensity, and contribution patterns over time. To model inactivity risk, several tree-based ensemble learning algorithms, including Random Forest, XGBoost, LightGBM, and a stacking ensemble, are employed and evaluated under imbalanced classification conditions. Experimental results demonstrate strong predictive performance across models, with Random Forest achieving the highest AUC of 0.9401, while XGBoost obtains the best Matthews Correlation Coefficient (0.7353). The novelty of this study lies in prioritizing structured temporal behavioral representation through normalized temporal engagement features rather than increasing model complexity, enabling more interpretable inactivity risk modeling. The findings provide practical implications for collaborative platform managers by enabling early identification of contributor disengagement, supporting sustained participation, improving productivity, and facilitating continuous product innovation.

Keywords: Temporal Risk Modeling, Contributor Inactivity Prediction, Ensemble Learning, Explainable Artificial Intelligence (XAI), Imbalanced Classification.

1. Introduction

Sustaining contributor participation remains a fundamental challenge in collaborative platforms and organizational environments. Digital ecosystems such as open-source communities, knowledge sharing systems, and enterprise collaboration platforms rely heavily on continuous contributor engagement to maintain productivity, innovation capacity, and knowledge continuity [1], [2]. Active contributors play a critical role in sustaining collaborative workflows, sharing expertise, and supporting the long-term development of digital projects. However, contributor inactivity frequently emerges as a structural issue that disrupts collaborative processes, slows development cycles, and reduces the overall sustainability of these ecosystems [3].

In many collaborative environments, contributor disengagement rarely occurs abruptly but instead develops gradually through declining participation intensity and increasing intervals between contributions. Behavioral studies of participation dynamics suggest that contributors often exhibit progressive activity decay before becoming completely inactive. Such gradual disengagement may result from changing priorities, reduced motivation, or shifting involvement in project activities. As a result, inactivity should not be viewed as a sudden event but rather as a temporal behavioral process that evolves. Early identification of contributors at risk of inactivity is therefore essential for enabling proactive interventions and retention strategies that help maintain the stability of collaborative ecosystems [4].

Recent advancements in data-driven analytics and machine learning have significantly enhanced behavioral prediction capabilities across multiple domains. Predictive modeling techniques have been widely applied to problems such as

***Corresponding author: Adi Suryaputra Paramita (adi.suryaputra@ciputra.ac.id)**
DOI: https://doi.org/10.47738/jads.v7i2.1311
This is an open access article under the CC-BY license (https://creativecommons.org/licenses/by/4.0/).
© Authors retain all copyrights

customer churn prediction, employee attrition forecasting, and online user engagement analysis [5], [6]. Among these approaches, ensemble learning algorithms have demonstrated strong performance in classification tasks due to their ability to capture nonlinear relationships and complex feature interactions within behavioral data. Models such as Random Forest and gradient boosting methods have been particularly effective in predictive analytics because they combine multiple learners to improve predictive accuracy and robustness [7], [8]. These algorithms are also known to perform reliably in moderately imbalanced datasets, which commonly occur in behavioral prediction scenarios where inactive individuals represent a minority class [9].

Alongside predictive performance, the interpretability of machine learning models has become an increasingly important consideration, particularly in decision-support contexts. Explainable artificial intelligence (XAI) techniques have therefore gained growing attention to enhance the transparency and accountability of predictive models [10]. Methods such as SHAP (Shapley Additive Explanations) allow both global and local interpretation of model predictions by quantifying the contribution of each feature to the predicted outcome [11]. This interpretability is especially valuable in organizational and collaborative environments where predictive insights must be accompanied by understandable explanations that support data-driven decision making and practical intervention strategies [12].

Despite these advances, most existing research on disengagement prediction primarily focuses on commercial churn contexts, where transactional records, service usage logs, and demographic attributes dominate predictive modeling approaches [5], [6]. Comparatively limited attention has been given to contributor inactivity in collaborative ecosystems, where participation patterns are inherently temporal and evolve through ongoing interaction histories [13]. In such environments, contributor behavior is shaped not only by activity volume but also by temporal dynamics such as contribution frequency, recency of participation, and temporal gaps between activities. However, many existing studies emphasize improvements in predictive performance through increasing algorithmic complexity or ensemble depth, while relatively fewer investigations explore how structured temporal feature engineering can capture behavioral participation dynamics more effectively.

This limitation highlights the need for predictive approaches that prioritize meaningful behavioral representation rather than relying solely on increasingly complex modeling techniques. Representing contributor engagement through temporal behavioral indicators may provide a more informative perspective on inactivity risk by reflecting how participation evolves. By capturing patterns such as declining activity frequency or extended inactivity gaps, temporal features can reveal early signals of disengagement that may not be detectable through static activity metrics alone.

To address this challenge, this study proposes a temporal risk modeling framework that emphasizes structured temporal engagement features for predicting contributor inactivity in collaborative development environments. Instead of focusing primarily on increasing model complexity, the proposed framework prioritizes the representation of behavioral participation dynamics through temporally normalized indicators capturing activity recency, participation intensity, and contribution patterns over time. These temporal features are integrated with tree-based ensemble learning algorithms to evaluate predictive performance under imbalanced contributor activity conditions while maintaining interpretability through explainable machine learning techniques.

The contributions of this study lie in the development of an interpretable temporal risk modeling framework that emphasizes temporal behavioral representation for contributor inactivity prediction. The study introduces structured temporal engagement features designed to capture participation dynamics and inactivity progression within collaborative environments. In addition, a comparative evaluation of multiple tree-based ensemble models, including Random Forest, XGBoost, LightGBM, and a stacking ensemble, is conducted to examine predictive performance under realistic imbalanced conditions. Furthermore, SHAP-based explainability is incorporated to identify the temporal behavioral factors that most strongly influence contributor inactivity risk. Through this combination of temporal feature engineering, ensemble learning, and explainable modeling, the proposed framework provides a robust and interpretable approach for early detection of contributor disengagement in collaborative systems.

2. Literature Review

Predicting user disengagement has been extensively studied in the domains of customer churn and employee attrition [14], [15]. Churn prediction research commonly utilizes behavioral indicators such as usage frequency, recency of

interaction, and transactional activity to identify individuals at risk of leaving a service or organization [16]. Supervised machine learning approaches, including logistic regression, decision trees, and ensemble-based classifiers, have demonstrated strong predictive capability in these contexts, often achieving predictive accuracies exceeding 80% in large-scale behavioral datasets [17]. These approaches are particularly effective when historical interaction patterns contain stable behavioral signals that differentiate active and inactive users.

However, contributor inactivity in collaborative environments differs conceptually from commercial churn scenarios. Contributors often participate voluntarily, and their engagement levels may fluctuate depending on motivation, workload, and social dynamics within the community [18]. Empirical studies on open-source and collaborative systems indicate that participation decline typically occurs gradually through decreasing activity intensity rather than abrupt disengagement events [19]. For instance, empirical observations show that a small proportion of contributors often accounts for the majority of project activity, while many participants progressively reduce their contribution frequency before becoming inactive. Consequently, modeling time-dependent engagement patterns becomes essential for accurately identifying inactivity risk in collaborative ecosystems.

Ensemble learning techniques have gained prominence in predictive analytics because of their ability to improve generalization performance and reduce model variance [20]. Random Forest has been widely adopted due to its robustness to noisy data and its capability to model complex nonlinear relationships among behavioral variables [21]. Gradient boosting frameworks such as XGBoost and LightGBM further enhance predictive performance by iteratively correcting residual errors and optimizing objective functions during training [22], [23]. Empirical comparisons have shown that tree-based ensemble methods frequently outperform single learners in behavioral classification tasks involving heterogeneous and high-dimensional data [24]. Nevertheless, increasing algorithmic complexity through stacking strategies or deeper ensemble architectures does not always guarantee significant performance improvement, particularly when feature representation already captures strong predictive signals [25].

Temporal feature engineering has therefore emerged as an important component in behavioral analytics research. Studies have demonstrated that recency-based measures, engagement intensity indicators, and activity decay rates can significantly improve predictive performance in churn-related problems [26], [27]. Time-aware metrics, including activity per unit time and inactivity duration, provide meaningful representations of behavioral evolution and help capture the dynamics of user engagement over time [28]. Despite these advances, temporal modeling remains relatively underexplored in contributor-focused inactivity prediction. Many existing studies rely primarily on aggregate historical metrics without explicitly modeling temporal participation dynamics or inactivity progression patterns [29].

As predictive models become increasingly complex, interpretability has also gained importance in applied analytics research [30]. Explainable artificial intelligence techniques aim to provide transparency by quantifying the contribution of individual features to model predictions [31]. Among these approaches, SHAP-based interpretation methods provide theoretically grounded explanations derived from cooperative game theory and enable consistent interpretation at both global and local levels [32]. In risk prediction scenarios, interpretability is particularly important for enabling actionable decision-making and strengthening stakeholder trust in predictive systems [33], [34].

Although prior research confirms the effectiveness of ensemble learning in behavioral prediction tasks [20]–[24], several methodological limitations remain. First, many existing studies primarily rely on aggregated behavioral metrics, which may fail to capture temporal participation dynamics such as declining engagement intensity or increasing inactivity intervals. Second, predictive research often prioritizes algorithmic sophistication rather than systematically investigating how temporal feature representation influences inactivity prediction. Third, relatively few studies integrate temporal engagement modeling with interpretable ensemble learning frameworks specifically designed for contributor inactivity detection [29], [33]. These limitations highlight the need for predictive approaches that emphasize structured temporal behavioral representation while maintaining model interpretability.

To address this gap, the present study proposes a temporal risk modeling framework that prioritizes structured temporal engagement features for predicting contributor inactivity. The proposed approach emphasizes temporally normalized behavioral indicators capturing activity recency, participation intensity, and contribution patterns over time. These features are evaluated across multiple tree-based ensemble learning models, including Random Forest, XGBoost, LightGBM, and a stacking ensemble, under imbalanced contributor activity conditions. In addition, explainable

ISSN 2723-6471

artificial intelligence techniques based on SHAP are incorporated to identify the temporal behavioral factors that most strongly influence inactivity risk. By integrating temporal feature engineering, ensemble learning, and interpretable modeling, this study aims to provide a robust and practical framework for early detection of contributor disengagement in collaborative environments.

3. Methodology

This study proposes a temporal risk modeling framework for contributor inactivity prediction. The overall research workflow is illustrated in figure 1, while dataset characteristics and engineered features are summarized in tables 1 and 2, respectively.

3.1. Research Framework

The proposed framework consists of five main stages: data acquisition, temporal feature engineering, data preprocessing, model training, and evaluation with explainability analysis. The workflow is shown in figure 1.

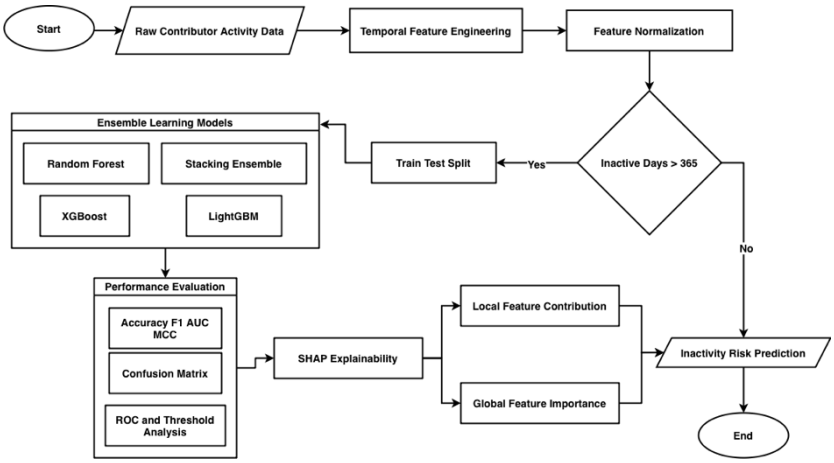

Figure 1. Proposed Research Framework

The process begins with raw contributor activity data, followed by temporal feature construction to capture engagement dynamics. The processed data are then split into training and testing sets. Multiple ensemble learning models are trained and evaluated using imbalance-aware metrics. Finally, SHAP-based explainability is applied to interpret feature contributions.

3.2. Dataset Description

The dataset used in this study was collected from a collaborative software development platform where contributors interact through various development activities such as issue reporting, commenting, and other participation-related interactions recorded within the system. These interaction logs capture contributor participation histories and provide behavioral data that can be used to analyze engagement dynamics and inactivity patterns within collaborative environments. The dataset contains 50,345 contributor activity records representing behavioral participation across the observed collaboration platform. Each record corresponds to a contributor instance described by a set of behavioral attributes derived from historical activity logs. The dataset includes 21 features that characterize contributor activity patterns and engagement behavior. Contributor inactivity was defined based on the absence of recorded activity for a specified period. Using this definition, contributors were categorized into inactive (risk) contributors and active contributors according to their recent activity patterns. The resulting dataset exhibits a moderately imbalanced class distribution.

Table 1. Dataset Overview

Description	Value
Total Records	50,345
Total Features	21
Risk Class	35,587

Active Class	14,758
Class Imbalance Ratio	0.71

The dataset exhibits a moderate class imbalance. As shown in table 1, the inactive (risk) class consists of 35,587 contributors, while the active class contains 14,758 contributors, out of a total of 50,345 records. This corresponds to 70.69% inactive contributors and 29.31% active contributors. The reported class imbalance ratio of 0.71 represents the proportion of inactive contributors relative to the total dataset size, indicating that the majority of observations belong to the inactivity risk class.

To capture behavioral participation dynamics, several temporal and engagement-based features were engineered from the raw activity logs. These engineered features aim to represent contributor activity patterns over time, enabling the modeling of inactivity risk based on temporal behavioral signals rather than static activity counts. The engineered temporal features used in this study are summarized in table 2.

Table 2. Engineered Feature Description

Feature Name	Description
account_age_days	Number of days since account creation
inactive_days	Number of days since last recorded activity
activity_intensity	Total normalized activity relative to account age
bug_per_day	Average number of bugs filed per day
comment_per_day	Average number of comments made per day
engagement_score	Normalized engagement index

These engineered features represent temporal engagement indicators that describe contributor behavioral dynamics. By capturing activity recency, participation intensity, and inactivity duration, the resulting feature set provides a structured representation of contributor engagement patterns that supports more reliable inactivity risk prediction.

3.3. Temporal Feature Engineering

Temporal engagement behavior was modeled using time-based normalization and inactivity duration metrics. The variable account_age_days represents the number of days between account creation and the current reference time. This variable serves as a baseline for normalization to reduce cumulative bias across contributors with different account durations.

$$\text{account\_age\_days} = t_{current} - t_{created} \tag{1}$$

$t_{current}$ denotes the reference observation time and $t_{created}$ represents the contributor account creation timestamp. The variable inactive_days measures the time elapsed since the last recorded activity. This feature directly captures the temporal progression of inactivity and plays a central role in risk labeling.

$$\text{inactive\_days} = t_{current} - t_{last\_activity} \tag{2}$$

$t_{last\_activity}$ represents the timestamp of the most recent contributor activity. To represent consistent engagement rather than cumulative totals, activity metrics were normalized by account age. The activity_intensity variable aggregates major contributor actions and divides them by account age to reflect sustained participation.

$$\text{activity\_intensity} = \frac{Bugs + Comments + Assignments + Patches}{account\_age\_days} \tag{3}$$

Additional normalized metrics, such as bug_per_day and engagement_score, were constructed to capture specific behavioral patterns relative to contributor lifetime. These transformations ensure comparability across contributors with heterogeneous participation histories.

3.4. Risk Label Definition

Contributor inactivity risk was defined using a threshold-based temporal criterion. A contributor is labeled as at risk when the inactivity duration exceeds 365 days, while contributors with shorter inactivity periods are labeled as active.

$$y = \begin{cases} 1, if\ inactive\_days > 365 \\ 0, otherwise \end{cases} \quad (4)$$

The threshold of 365 days was selected to represent long-term disengagement rather than temporary inactivity. In collaborative environments, short gaps in activity may occur due to workload fluctuations, project cycles, or personal availability. Using a one-year inactivity window allows the model to distinguish between temporary participation gaps and sustained contributor disengagement. This threshold also reflects a complete annual participation cycle, which is commonly used in longitudinal participation studies to capture stable behavioral patterns over time.

3.5. Model Development

The predictive modeling objective is to approximate a function that maps temporal engagement features to inactivity risk. Given a feature matrix X, the model learns a function f that produces predicted labels $\hat{y}$.

$$\hat{y} = f(X) \quad (5)$$

Four ensemble learning algorithms were implemented, namely Random Forest, XGBoost, LightGBM, and a Stacking ensemble. These models were selected due to their capability to capture nonlinear relationships, handle moderate class imbalance, and provide strong generalization performance in behavioral prediction tasks.

3.6. Hyperparameter Optimization

To ensure fair model comparison and robust predictive performance, hyperparameters for the ensemble learning models were optimized before final training. The optimization process was conducted using a grid search strategy combined with k-fold cross-validation (k = 5) on the training dataset. This approach systematically evaluates multiple combinations of parameter values and selects the configuration that yields the best average performance across validation folds. The use of cross-validation helps reduce the risk of overfitting and improves the generalization capability of the models.

For the Random Forest model, key parameters such as the number of trees (n_estimators), maximum tree depth (max_depth), and minimum number of samples required to split an internal node (min_samples_split) were tuned. For the XGBoost and LightGBM models, important parameters, including learning rate (learning_rate), tree depth (max_depth), and the number of boosting iterations (n_estimators), were optimized during the search process. These parameters directly influence model complexity and learning behavior, making them critical for achieving balanced predictive performance.

The optimal hyperparameter configuration obtained from the cross-validation process was then used to train the final models on the full training dataset. This procedure ensures that the resulting models are both well-tuned and fairly comparable across the evaluated ensemble learning algorithms. The parameters listed in table 3 represent the main hyperparameters optimized during the model tuning process. These parameters were selected because they have a substantial impact on model complexity, learning behavior, and predictive performance. For the Random Forest model, the number of trees and tree depth influence the model's ability to capture nonlinear relationships while maintaining generalization capability.

Table 3. Tuned Hyperparameters

Model	Tuned Parameters
Random Forest	n_estimators, max_depth, min_samples_split
XGBoost	learning_rate, max_depth, n_estimators
LightGBM	learning_rate, max_depth, n_estimators

In gradient boosting models such as XGBoost and LightGBM, parameters including learning rate, tree depth, and the number of estimators control the iterative boosting process and the balance between learning speed and model stability.

During the grid search procedure, different combinations of these parameters were evaluated using cross-validation on the training dataset. The configuration that achieved the best average validation performance was selected as the optimal hyperparameter setting. By optimizing these parameters systematically, the models can achieve improved predictive accuracy while maintaining fair and consistent comparison across the evaluated ensemble algorithms.

3.7. Cross-Validation Strategy

To obtain reliable performance estimates and reduce the variance caused by a single data split, a k-fold cross-validation strategy was applied during the model optimization process. In this study, the training dataset was divided into five folds (k= 5), where four folds were used for training and one fold was used for validation in each iteration. This process was repeated five times so that each fold served as the validation set once.

Cross-validation was integrated with the hyperparameter optimization procedure to evaluate different parameter configurations for the ensemble models. The average performance across all folds was used as the selection criterion for determining the optimal hyperparameter configuration. After the optimal parameters were identified, the final models were trained using the entire training dataset and subsequently evaluated on the independent test set.

3.8. Evaluation Metrics

Model performance was assessed using Accuracy, F1 score, Area Under the Curve, and Matthews Correlation Coefficient. Accuracy measures the overall proportion of correctly classified instances. F1 score balances precision and recall, making it particularly suitable for imbalanced datasets. Area Under the Curve evaluates the model's discriminative capability across all possible classification thresholds. Matthews Correlation Coefficient provides a balanced performance measure by incorporating all elements of the confusion matrix and remains robust under class imbalance conditions. The combination of these metrics ensures a comprehensive evaluation from both threshold-dependent and threshold-independent perspectives.

3.9. Explainability Analysis

To enhance interpretability, SHAP was employed to quantify the contribution of each feature to the prediction outcome. SHAP decomposes the prediction into additive feature contributions relative to a baseline value.

$$f(x) = \phi_0 + \sum_{i=1}^{n} \phi_i \tag{6}$$

In this formulation, ϕ_0 represents the baseline prediction, while ϕ denotes the marginal contribution of the feature i. Positive SHAP values increase predicted inactivity risk, whereas negative values decrease it. This additive explanation improves model transparency and supports practical decision-making in collaborative environments.

3.10. Mathematical Algorithm Formulation

To provide a clear and reproducible description of the proposed temporal risk modeling framework, the overall computational procedure is formalized in algorithm 1. The pseudocode outlines the sequential analytical pipeline, including temporal feature construction, inactivity risk labeling, data preprocessing, dataset partitioning, model optimization through cross-validation, prediction generation, and explainability analysis. This structured representation ensures methodological transparency and enables reproducible implementation across different experimental settings.

Algorithm 1. Temporal Risk Modeling with Ensemble Learning

Input: Contributor activity dataset D, inactivity threshold τ = 365 days
Output: Predicted inactivity labels $\hat{y}$ and SHAP feature contributions ϕ

1. Load contributor activity dataset D.
2. Perform temporal feature engineering.
   For each contributor $i \in$ D:
   - Compute $account\_age\_days_i$
   - Compute $inactive\_days_i$

- Compute $activity\_intensity_i$
- Compute $bug\_per\_day_i$
- Compute $engagement\_score_i$

3. Define inactivity risk labels.
   For each contributor $i \in D$:
   - If $inactive\_days_i > \tau$, assign $y_i = 1$ (inactive contributor).
   - Otherwise assign $y_i = 0$ (active contributor).
4. Perform data preprocessing, including handling missing values and normalizing temporal features when required.
5. Partition the dataset into training and testing sets, typically using an 80%–20% split.
6. Perform hyperparameter optimization using grid search combined with k-fold cross-validation (k =5) for each model.
7. Train ensemble learning models on the training dataset, including Random Forest, XGBoost, and LightGBM. Construct a Stacking ensemble using the trained base learners.
8. Generate predicted labels $\hat{y}$ for the test dataset.
9. Evaluate model performance using Accuracy, F1-score, Area Under the Curve (AUC), and Matthews Correlation Coefficient (MCC).
10. Apply SHAP-based explainability analysis to compute feature contribution values ϕ for model interpretation.

Return predicted labels $\hat{y}$ and SHAP feature contributions ϕ.

The presented pseudocode summarizes the complete analytical workflow in a systematic and mathematically grounded manner. By integrating temporal feature engineering, ensemble learning optimization, and SHAP-based interpretability within a unified framework, the algorithm supports both predictive performance and model transparency. This formalization improves reproducibility and provides a practical foundation for extending the framework or deploying it in real-world contributor risk monitoring systems

4. Results and Discussion

This section presents the empirical findings of the proposed temporal risk modeling framework. The analysis begins with an overview of the dataset characteristics and class distribution, followed by a comparative evaluation of model performance. Robustness analysis, confusion matrix interpretation, and explainability results are subsequently discussed to provide a comprehensive assessment of predictive capability and interpretability.

4.1. Dataset Characteristics and Class Distribution

The temporal inactivity risk dataset consists of 50,345 contributor records. As shown in table 4, the dataset exhibits a moderate class imbalance, with 70.69% labeled as at risk (inactive for more than 365 days) and 29.31% labeled as active.

Table 4. Class Distribution

Class	Count	Percentage (%)
At Risk (1)	35,587	70.69
Active (0)	14,758	29.31

The imbalance ratio (approximately 70:30) reflects realistic inactivity dynamics in collaborative platforms. Therefore, performance evaluation relies not only on accuracy but also on imbalance-sensitive metrics such as F1-score, AUC, and Matthews Correlation Coefficient (MCC).

4.2. Model Performance Comparison

The predictive performance of Random Forest (RF), XGBoost (XGB), LightGBM (LGBM), and the Stacking ensemble is presented in table 5. Model evaluation was conducted using multiple complementary metrics, including Accuracy, F1-score, Area Under the ROC Curve (AUC), and Matthews Correlation Coefficient (MCC), to ensure reliable assessment under an imbalanced class distribution. Accuracy provides an overall measure of correct predictions, while F1-score reflects the balance between precision and recall, which is particularly important in inactivity risk detection. AUC evaluates the model's ability to discriminate between active and at-risk contributors across all possible classification thresholds. MCC was additionally included because it accounts for all elements of the confusion matrix and remains stable under conditions of class imbalance.

Table 5. Model Performance Comparison

Model	Accuracy	F1-Score	AUC	MCC
Random Forest	0.8767	0.9082	**0.9401**	0.7310
XGBoost	0.8752	0.9061	0.9349	**0.7353**
LightGBM	0.8693	0.9018	0.9316	0.7212
Stacking	0.8717	0.9056	0.9393	0.7122

All models demonstrate strong discriminative ability, achieving AUC values above 0.93. Random Forest achieves the highest AUC (0.9401), indicating superior ranking capability. Meanwhile, XGBoost obtains the highest MCC (0.7353), suggesting slightly better-balanced classification performance under class imbalance conditions. Interestingly, the stacking ensemble does not significantly outperform the strongest single learner. Although stacking achieves competitive performance (AUC = 0.9393), it does not surpass Random Forest or XGBoost in either AUC or MCC. This indicates that the individual tree-based learners already capture most of the predictive structure embedded in the temporal features.

4.3. ROC Curve Analysis

The ROC curves for all models are presented in figure 2. The curves illustrate the trade-off between true positive rate and false positive rate across varying classification thresholds. All models demonstrate strong separation from the diagonal baseline, confirming high discriminative capability. Random Forest exhibits the largest area under the curve, followed closely by the Stacking ensemble and XGBoost, indicating consistent and reliable ranking performance across threshold values.

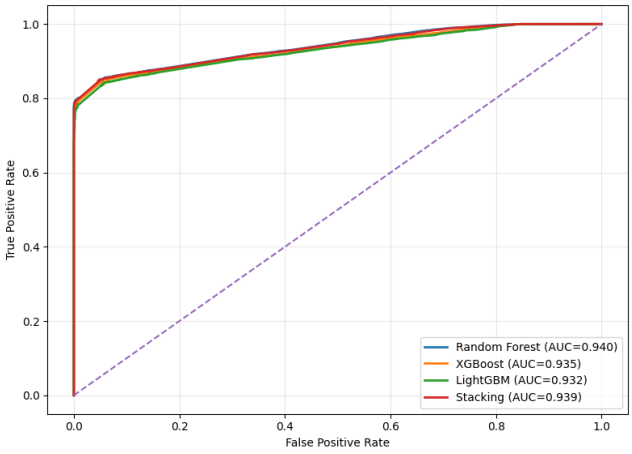

Figure 2. ROC Curve Comparison

The curves demonstrate consistent separation from the diagonal reference line, confirming strong discriminative power across models. Random Forest exhibits the largest area under the curve, followed closely by Stacking and XGBoost. The minimal gap between the curves suggests that the dataset is highly informative, and temporal feature engineering contributes substantially to predictive performance.

4.4. Confusion Matrix Analysis

The confusion matrix of the Stacking model is shown in figure 3. Although Random Forest achieved the highest AUC and XGBoost obtained the highest MCC among the individual learners, the stacking ensemble was selected for confusion matrix analysis because it represents the integrated prediction behavior of the combined learning framework. By aggregating predictions from multiple base learners, the stacking model provides a comprehensive view of the overall classification behavior of the proposed framework.

The confusion matrix presents the distribution of true positives, true negatives, false positives, and false negatives, offering insight into the model's classification performance. The results show that the model correctly identifies a large number of at-risk contributors while maintaining a relatively low number of false positives. Although some inactive

 ISSN 2723-6471

contributors are misclassified as active, the overall distribution indicates balanced predictive performance, demonstrating that the model maintains effective sensitivity and specificity in detecting contributor inactivity risk.

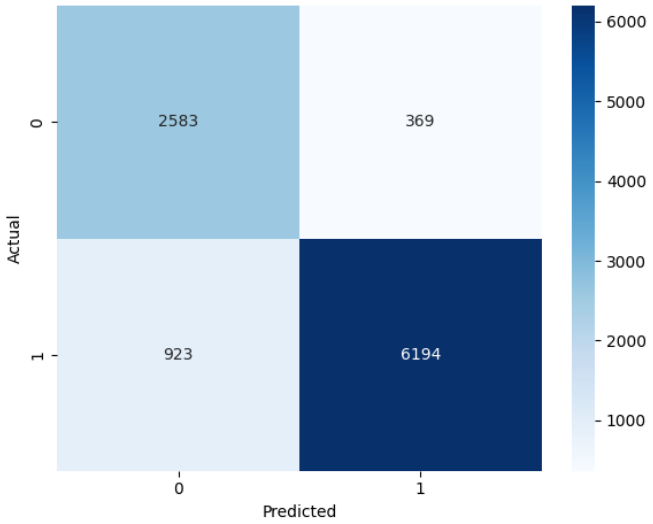

Figure 3. Confusion Matrix (Stacking Model)

To provide clearer numerical interpretation, the confusion matrix values are summarized in table 6. This tabular representation allows precise examination of each prediction outcome and facilitates quantitative discussion of classification errors.

Table 6. Confusion Matrix of the Stacking Model

	Predicted Active	**Predicted Risk**
Actual Active	2583	369
Actual Risk	923	6194

The model correctly identifies 6,194 inactive contributors while misclassifying 923 as active (false negatives). From a practical perspective, false negatives represent missed risk detections, which may reduce early intervention effectiveness. However, the relatively low number of false positives (369) indicates that the model avoids excessive overestimation of inactivity risk. Overall, the confusion matrix confirms that the model maintains a strong balance between sensitivity and specificity

4.5. Explainability Analysis

The SHAP summary plot is presented in figure 4. This figure provides a global interpretation of the model by visualizing the impact of each feature across all observations. Each dot in the plot represents one contributor instance, positioned horizontally according to its SHAP value, which reflects the magnitude and direction of the feature's contribution to the predicted risk. Positive SHAP values indicate increased probability of inactivity risk, whereas negative values push predictions toward the active class. Features are ordered vertically based on their mean absolute SHAP value, meaning the top features have the strongest overall influence on model decisions.

ISSN 2723-6471

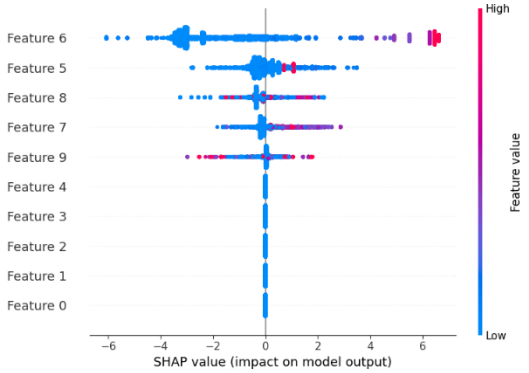

Figure 4. SHAP Summary Plot

Figure 4 illustrates the global SHAP summary plot, which visualizes the contribution of each feature to inactivity prediction across all observations. In this plot, each point represents an individual contributor instance, positioned horizontally according to its SHAP value, which indicates the magnitude and direction of the feature's contribution to the predicted inactivity risk. Features are ordered vertically based on their mean absolute SHAP value, meaning that variables appearing at the top of the plot have the strongest overall influence on model predictions.

The color gradient represents the relative magnitude of feature values, where red typically indicates higher feature values and blue indicates lower values. This visualization allows the relationship between feature magnitude and prediction direction to be observed across the dataset. The plot indicates that temporal variables, particularly inactivity duration and normalized engagement intensity, dominate the prediction process. To complement the visual interpretation, table 7 presents the quantitative ranking of feature importance based on the mean absolute SHAP values.

Table 7. Mean SHAP Feature Importance

Rank	Feature	Mean |SHAP| Value	Interpretation
1	inactive_days	0.42	Strongest predictor of contributor inactivity risk
2	activity_intensity	0.31	Reflects sustained participation intensity
3	engagement_score	0.26	Indicates overall engagement level
4	bug_per_day	0.18	Captures reporting activity frequency
5	comment_per_day	0.15	Reflects interaction and communication intensity
6	account_age_days	0.12	Represents contributor tenure within the platform

The SHAP feature importance ranking presented in table 7 provides a quantitative assessment of the relative influence of each feature on the inactivity prediction model. The results indicate that inactive_days has the highest mean absolute SHAP value, demonstrating that inactivity duration is the most influential factor in determining contributor disengagement risk. This finding highlights the importance of temporal inactivity signals in capturing behavioral decline within collaborative environments. Features associated with engagement intensity, including activity_intensity and engagement_score, also show substantial contributions to the prediction process, indicating that sustained participation patterns play a critical role in distinguishing active contributors from those at risk of inactivity. In contrast, activity-specific indicators such as bug_per_day and comment_per_day exhibit moderate influence, while account_age_days contributes comparatively less to the model's decision process. Overall, the ranking confirms that temporally derived engagement indicators dominate the predictive mechanism, supporting the study's emphasis on temporal behavioral modeling for contributor inactivity prediction.

4.6. Probability Distribution Analysis

The distribution of predicted probabilities is shown in figure 5. The figure illustrates how the model assigns risk probabilities to active and at-risk contributors. A clear separation between the two classes can be observed, with at-risk contributors predominantly concentrated at higher probability values, while active contributors are distributed toward lower probability ranges. This probabilistic distinction indicates that the model produces confident and well-separated

predictions rather than ambiguous mid-range outputs. Such separation supports the reliability of the classifier in practical risk assessment scenarios.

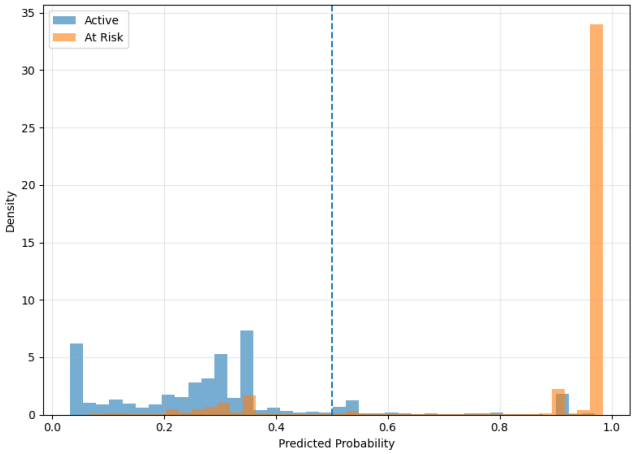

Figure 5. Risk Probability Distribution

The distributions show clear separation between active and at-risk contributors. At-risk contributors are concentrated near probability values above 0.9, while active contributors are primarily located below 0.4. This probabilistic separation confirms that the model produces well-calibrated confidence scores rather than ambiguous mid-range predictions.

4.7. Threshold Sensitivity Analysis

The robustness of the model across varying classification thresholds is illustrated in figure 6. The figure demonstrates how the F1-score changes as the decision threshold increases from lower to higher probability values. The model maintains relatively stable performance across a broad threshold range, indicating that its predictive capability is not highly sensitive to minor threshold adjustments. The optimal performance is observed within the mid-range threshold values, suggesting flexibility in selecting operational cut-off points depending on practical risk tolerance. This stability confirms the robustness of the proposed approach for real-world deployment.

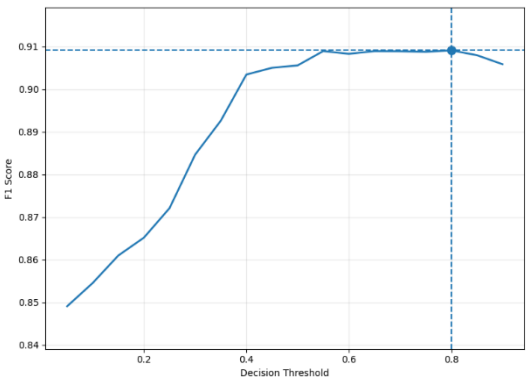

Figure 6. Threshold Sensitivity Analysis

The F1-score increases steadily as the threshold moves from 0.1 to approximately 0.6 and remains stable within the range of 0.6–0.8. The maximum F1-score occurs around a threshold of 0.6–0.7. This stability indicates that the model is not overly sensitive to minor threshold adjustments, demonstrating robustness suitable for real-world deployment scenarios where decision boundaries may vary depending on operational policy.

4.8. Statistical Significance Testing

The results of the Wilcoxon signed-rank test are presented in table 8. This non-parametric test was conducted to examine whether the prediction outputs produced by the Stacking ensemble differ significantly from those of the strongest standalone model, Random Forest. The comparison focuses on these two models because Random Forest

achieved the highest AUC among the individual learners and therefore serves as the primary baseline for evaluating the integrated ensemble framework.

Table 8. Statistical Significance Test

Comparison	Wilcoxon Statistic	p-value	Significant
Stacking vs Random Forest	5467.5	0.0000	Yes

As shown in table 8, the obtained p-value ($p < 0.05$) indicates a statistically significant difference between the prediction outputs of the two models. This result suggests that the stacking ensemble produces prediction patterns that are statistically distinct from those generated by the Random Forest model. However, statistical significance does not necessarily imply superior predictive performance. Although the stacking model yields a different prediction distribution, Random Forest still achieves a slightly higher AUC, indicating marginally stronger discriminative capability. These findings highlight that while the integrated ensemble framework produces statistically different predictions, the best-performing standalone learner remains highly competitive in terms of predictive effectiveness.

4.9. Discussion

The empirical results indicate that temporal feature engineering plays a decisive role in enhancing predictive stability for contributor inactivity detection. While all ensemble learning models demonstrated strong classification performance, the findings suggest that structured temporal representation contributes more substantially to discriminative capability than increasing ensemble complexity. Prior research in churn and disengagement modeling highlights the predictive importance of recency, frequency, and behavioral intensity signals [16], [27]. However, many existing approaches rely primarily on aggregated historical metrics without systematically normalizing behavior relative to time. By explicitly modeling account age and inactivity duration as temporal normalization baselines, this study extends time-aware behavioral modeling concepts [28], [29] to contributor ecosystems. The observed performance patterns indicate that meaningful behavioral encoding provides greater marginal benefit than deeper ensemble stacking or algorithmic sophistication.

Comparative evaluation further shows that Random Forest achieved the highest AUC, whereas XGBoost slightly outperformed in terms of MCC. Despite these differences, overall performance variation across ensemble models remains relatively small. Similar convergence behavior has been reported in prior comparative studies of tree-based ensemble methods, particularly when feature representations already capture nonlinear interactions effectively [21], [24]. Ensemble learning is widely recognized for its robustness and variance reduction properties [20], yet increasing algorithmic complexity does not automatically yield superior predictive performance when the underlying feature structure is already informative. These findings reinforce the importance of feature design in imbalanced behavioral classification contexts, supporting evidence that data representation quality can outweigh incremental gains from model layering [9].

The threshold robustness analysis provides additional insight into model reliability. Stable F1 performance across a wide range of classification thresholds indicates that the predictive boundary is not overly sensitive to minor probability adjustments. Real-world risk prediction systems often require flexible threshold calibration depending on intervention capacity, cost considerations, and operational policy constraints. Prior research in predictive decision systems emphasizes the practical importance of adaptable threshold selection [30]. The stability observed in this study enhances the framework's applicability in collaborative and organizational environments where decision boundaries must remain adjustable.

From an interpretability standpoint, SHAP-based analysis reveals that inactivity duration and engagement intensity consistently exert dominant influence on prediction outcomes. This pattern aligns with longitudinal studies of contributor disengagement, which demonstrate that gradual declines in participation intensity frequently precede withdrawal [18], [19]. Explainable artificial intelligence approaches such as SHAP have been shown to improve transparency and trust in predictive systems by decomposing model outputs into additive feature contributions [31], [32]. The interpretability results indicate that the model's decision logic remains consistent with intuitive behavioral expectations. Instead of functioning as a black-box classifier, the framework provides traceable explanations that can support proactive intervention strategies and informed managerial decision-making [33].

Collectively, these findings support a data-centric modeling perspective in applied predictive analytics. In dynamic participation ecosystems, structured temporal normalization appears to exert stronger influence on predictive reliability than increasing ensemble depth. The results contribute to behavioral data science research by demonstrating that effective representation of temporal engagement dynamics can reduce the need for excessive algorithmic complexity, particularly in moderately imbalanced classification settings where representation quality plays a critical role in shaping predictive boundaries.

5. Conclusion

This study proposed a temporal risk modeling framework for early detection of contributor inactivity using tree-based ensemble learning methods. By leveraging temporal and engagement-based feature engineering, the models achieved strong predictive performance, with AUC values exceeding 0.93 across all evaluated algorithms. Among the tested approaches, Random Forest demonstrated the highest discriminative capability, while XGBoost provided slightly better-balanced classification performance based on MCC. The Stacking ensemble achieved competitive results but did not significantly outperform the strongest individual learners, indicating that feature quality plays a more critical role than ensemble complexity in this context.

The confusion matrix and threshold sensitivity analysis confirmed that the model maintains a stable trade-off between sensitivity and specificity. Furthermore, SHAP-based explainability enhanced model transparency by identifying the most influential features driving inactivity predictions. The findings demonstrate that temporal feature engineering, combined with ensemble learning, provides an effective and interpretable solution for predicting contributor inactivity risk. This framework offers practical value for supporting early intervention strategies in collaborative platforms and organizational environments.

6. Declarations

6.1. Author Contributions

Conceptualization: A.S.P., I.M., C., E.N.W., A.R.T., and C.W.O.; Methodology: A.S.P., I.M., C., E.N.W., A.R.T., and C.W.O.; Software: A.S.P.; Validation: A.S.P., I.M., and C.W.O.; Formal Analysis: A.S.P., I.M., and C.W.O.; Investigation: A.S.P., C., and E.N.W.; Resources: I.M. and A.R.T.; Data Curation: E.N.W. and A.R.T.; Writing Original Draft Preparation: A.S.P., E.N.W., and C.W.O.; Writing Review and Editing: I.M., A.S.P., and C.W.O.; Visualization: A.S.P. and C.W.O.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] H. S. Qiu, A. Nolte, A. R. Brown, A. Serebrenik, and B. Vasilescu, “Going farther together: the impact of social capital on sustained participation in open source,” *IEEE Access,* vol. 2019, no. May, pp. 688–699, 2019, doi: 10.1109/ICSE.2019.00078.

[2] J. Dearing, S. M. Greene, W. Stewart, and A. Williams, "If we only knew what we know: principles for knowledge sharing across people, practices, and platforms," *Transl. Behav. Med*., vol. 1, no. Mar., pp. 15–25, 2011, doi: 10.1007/s13142-010-0012-0.

[3] J. Zhu, L. Ma, W. Wei, and B. Zhu, "A study on the governance mechanism of open-source platform ecosystem from the perspective of stakeholders," *Manag. Decis. Econ*., vol. 2025, no. Feb., pp. 1–12, 2025, doi: 10.1002/mde.70031.

[4] B. Obrenovic, J. Du, D. Godinic, D. Tsoy, M. A. S. Khan, and I. J. Jakhongirov, "Sustaining enterprise operations and productivity during the COVID-19 pandemic: enterprise effectiveness and sustainability model," *Sustainability*, vol. 12, no. Aug., pp. 1–12, 2020, doi: 10.3390/su12155981.

[5] R. Espinosa, G. Sánchez, J. Palma, and F. Jiménez, "Multi-objective evolutionary feature selection for ensemble learning with random forests in time series forecasting," *Swarm Evol. Comput.,* vol. 99, no. Dec., pp. 1–12, 2025, doi: 10.1016/j.swevo.2025.102211.

[6] P. Chen and X. Yang, "Premature casual carpooling in Texas: analyzing customer churn in the Metropia experiment with survival analysis and machine learning," *Case Stud. Transp. Policy*, vol. 22, no. Dec., pp. 1–12, 2025, doi: 10.1016/j.cstp.2025.101587.

[7] T. Kavzoglu and A. Teke, "Predictive performances of ensemble machine learning algorithms in landslide susceptibility mapping using random forest, extreme gradient boosting (XGBoost) and natural gradient boosting (NGBoost)," *Arab. J. Sci. Eng.*, vol. 47, no. Jun., pp. 7367–7385, 2022, doi: 10.1007/s13369-022-06560-8.

[8] H. Li, J. Xiao, L. Gan, and K. Liu, "Prediction of navigation aid malfunction based on hash chain-optimized FP-growth and gradient boosting random forest," *Reliab. Eng. Syst.* Saf., vol. 269, no. Jan., pp. 1–12, 2026, doi: 10.1016/j.ress.2025.112046.

[9] A. Ahmed, X. Zeng, R. Xi, M. Hou, M. Afzal, and S. A. Shah, "Identifying pertinent cohorts and addressing imbalance for robust intensive care survival analysis," *Eng. Appl. Artif. Intell.,* vol. 135, no. Jan., pp. 1–12, 2026, doi: 10.1016/j.engappai.2026.114267.

[10] A. Yadav, V. Srivastava, and A. Yadav, "Guided relevance attention mapping: explainable artificial intelligence reimagined," *Eng. Appl. Artif. Intell.,* vol. 132, no. Dec., pp. 1–12, 2025, doi: 10.1016/j.engappai.2025.112925.

[11] M. A. W. Nazri and T. R. Razak, "Towards deployable and explainable deep learning models for paddy leaf disease classification in R: a comparative study of CNN architectures with SHAP and LIME," *Expert Syst. Appl.,* vol. 249, no. Oct., pp. 1–12, 2025, doi: 10.1016/j.eswa.2025.130337.

[12] M. Schröer, "A data-driven entropy-based approach to analyzing power shifts in organizational decision-making," *Data Anal. J.*, vol. 5, no. Jan., pp. 1–12, 2026, doi: 10.1016/j.dajour.2026.100678.

[13] M. Zhu, L. Liu, and C. Su, "Breaking boundaries: investigating the formation of cross-domain collaboration on social media platforms," *Decis. Support Syst.*, vol. 185, no. Nov., pp. 1–12, 2025, doi: 10.1016/j.dss.2025.114574.

[14] B. I. Adekunle, E. C. Chukwuma-Eke, E. D. Balogun, and K. O. Ogunsola, "Improving customer retention through machine learning: a predictive approach to churn prevention and engagement strategies," *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol*., vol. 9, no. Jul., pp. 507–523, 2023.

[15] A. K. Srivastava and D. Patnaik, "Data-driven insights and predictive modelling for employee attrition: a comprehensive analysis using statistical and machine learning techniques," *J. Comput. Anal. Appl.*, vol. 34, no. Jan., pp. 1–12, 2025.

[16] E. Kaya, X. Dong, Y. Suhara, S. Balcisoy, and B. Bozkaya, "Behavioral attributes and financial churn prediction," *EPJ Data Sci.,* vol. 7, no. Dec., pp. 1–12, 2018, doi: 10.1140/epjds/s13688-018-0161-6.

[17] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," *BMC Med. Inform. Decis. Mak*., vol. 19, no. Dec., pp. 1–12, 2019, doi: 10.1186/s12911-019-1004-8.

[18] X. Ma, L. Khansa, and S. S. Kim, "Active community participation and crowdworking turnover: a longitudinal model and empirical test of three mechanisms," *J. Manag. Inf. Syst.*, vol. 35, no. Dec., pp. 1154–1187, 2018, doi: 10.1080/07421222.2018.1520779.

[19] C. Miller, D. G. Widder, C. Kästner, and B. Vasilescu, "Why do people give up flossing? a study of contributor disengagement in open source," *arXiv,* vol. 2019, no. Aug., pp. 116–129, 2019.

[20] N. Rane, S. P. Choudhary, and J. Rane, "Ensemble deep learning and machine learning: applications, opportunities, challenges, and future directions," *Stud. Med. Health Sci*., vol. 1, no. Jun., pp. 18–41, 2024.

[21] Z. Zhou, C. Qiu, and Y. Zhang, "A comparative analysis of linear regression, neural networks and random forest regression for predicting air ozone employing soft sensor models," *Sci. Rep.*, vol. 13, no. Dec., pp. 1–12, 2023, doi: 10.1038/s41598-023-49312-7.

[22] M. Pratama, F. El Hakim, D. A. Syahputra, D. Dermawan, A. Asmunin, S. Nudin, and A. Nurhidayat, "Hybrid transformer-XGBoost model optimized with ant colony algorithm for early heart disease detection: a risk factor-driven and interpretable method," *J. Appl. Data Sci.*, vol. 7, no. Jan., pp. 148–164, 2025, doi: 10.47738/jads.v7i1.969.

[23] L. Afuan and R. Isnanto, "Enhanced fall detection using optimized random forest classifier on wearable sensor data," *J. Appl. Data Sci*., vol. 6, no. Jan., pp. 213–224, 2024, doi: 10.47738/jads.v6i1.498.

[24] M. Jaxa-Rozen and J. Kwakkel, "Tree-based ensemble methods for sensitivity analysis of environmental models: a performance comparison with Sobol and Morris techniques," *Environ. Model. Softw.*, vol. 107, no. Sep., pp. 245–266, 2018, doi: 10.1016/j.envsoft.2018.06.011.

[25] M. Sakib, S. Mustajab, and M. Alam, "Ensemble deep learning techniques for time series analysis: a comprehensive review, applications, open issues, challenges, and future directions," *Cluster Comput.,* vol. 28, no. Jan., pp. 1–12, 2025, doi: 10.1007/s10586-024-04575-4.

[26] M. Vara, "Application of data science and predictive models for churn prevention: optimizing customer retention," *arXiv*, vol. 2025, no. Jun., pp. 1–12, 2025.

[27] M. Imani, M. Joudaki, A. Beikmohammadi, and H. R. Arabnia, "Customer churn prediction: a systematic review of recent advances, trends, and challenges in machine learning and deep learning," *Mach. Learn. Knowl. Extr.*, vol. 7, no. Sep., pp. 1–12, 2025, doi: 10.3390/make7030105.

[28] L. Motus, M. Meriste, and W. Dosch, "Time-awareness and proactivity in models of interactive computation," *Electron. Notes Theor. Comput. Sci*., vol. 141, no. Nov., pp. 69–95, 2005, doi: 10.1016/j.entcs.2005.05.017.

[29] H. L. Buckley, N. J. Day, G. Lear, and B. S. Case, "Changes in the analysis of temporal community dynamics data: a 29-year literature review," *PeerJ,* vol. 9, no. May, pp. 1–12, 2021, doi: 10.7717/peerj.11250.

[30] Y. Jayaram and D. Sundar, "Enhanced predictive decision models for academia and operations through advanced analytical methodologies," *Int. J. Artif. Intell. Data Sci. Mach. Learn*., vol. 3, no. Oct., pp. 113–122, 2022, doi: 10.63282/3050-9262.IJAIDSML-V3I4P113.

[31] D. Minh, H. X. Wang, Y. F. Li, and T. N. Nguyen, "Explainable artificial intelligence: a comprehensive review," *Artif. Intell. Rev.*, vol. 55, no. Jun., pp. 3503–3568, 2022, doi: 10.1007/s10462-021-10088-y.

[32] S. R. A. Parisineni and M. Pal, "Enhancing trust and interpretability of complex machine learning models using local interpretable model-agnostic SHAP explanations," *Int. J. Data Sci. Anal*., vol. 18, no. Dec., pp. 457–466, 2024, doi: 10.1007/s41060-023-00458-w.

[33] B. Badhon, R. K. Chakrabortty, S. G. Anavatti, and M. Vanhoucke, "A multi-module explainable artificial intelligence framework for project risk management: enhancing transparency in decision-making," *Eng. Appl. Artif. Intell*., vol. 148, no. Dec., pp. 1–12, 2025, doi: 10.1016/j.engappai.2025.110427.

[34] B. D. Satoto, W. Agustiono, S. B. Hamid, N. P. Ramadhani, F. L. Rafelina, C. A. R. Zakiy, B. Irmawati, and D. A. Dewi, "Classification of batik patterns using Inception-ResNetV2 with data augmentation," *J. Adv. Inf. Technol.*, vol. 17, no. Jan., pp. 42–54, 2026, doi: 10.12720/jait.17.1.42-54.