Journal of Artificial Intelligence in Information Technology (JAI-IT)
Vol. 1 No. 1 June 2026
E-ISSN: XXXX-XXXX; P-ISSN: XXXX-XXXX, Pages XXX-XXX
DOI: https://doi.org/10.51903/7d0szq86

A Comparative Evaluation of Machine Learning Models in Enterprise Information Systems with SHAP-Based Explainability Analysis

Azka Nahya Amanta*[1,2], Lukman Santoso[2]
Email: azkanahya0035@mhs.unisbank.ac.id (1), lukman@stekom.ac.id (2)
Orcid: https://orcid.org/0009-0004-1709-6384 (1), https://orcid.org/0009-0004-1709-6384 (2)
[1]*Department of Information Technology. Faculty of Information Technology and Industry. Universitas Stikubank, Semarang, Indonesia 50241*
[2]*Department of Computer System. Faculty of Academic Studies. Universitas Sains dan Teknologi Komputer, Semarang, Indonesia 50192*
*Corresponding Author

Abstract

Enterprise Information Systems (EIS) generate large volumes of transactional data across functional domains including Human Resource Management (HRM), Customer Relationship Management (CRM), Supply Chain Management (SCM), and Financial Management. Although Decision Tree (DT), Random Forest (RF), and Support Vector Machine (SVM) are among the most widely applied classification algorithms, no study has systematically benchmarked these three classifiers across multiple enterprise IS domains under consistent methodological conditions. This study presents a comparative evaluation of DT, RF, and SVM applied to four representative enterprise IS classification tasks: employee attrition, customer churn, supplier delay risk, and financial fraud detection. The experimental framework incorporates the Synthetic Minority Oversampling Technique (SMOTE) to address class imbalance endemic to enterprise data. It employs grid search cross-validation for hyperparameter optimization to ensure fair comparison. SHAP-based explainability analysis is further applied to the Random Forest (RF) classifier across all four domains bridging algorithmic performance and enterprise decision-makers' interpretability. Results show that RF consistently achieves the highest predictive performance across all four domains, while SVM demonstrates strong stability, and DT retains advantages in interpretability and computational efficiency. The study culminates in a practitioner-oriented algorithm selection framework that guides enterprise IS stakeholders in choosing appropriate classifiers based on domain characteristics.

Keywords: *classification, data mining, decision tree, enterprise information systems, explainability, random forest, SHAP, SMOTE, support vector machine.*

I. INTRODUCTION

The rapid proliferation of digital transformation has fundamentally reshaped the operational landscape of modern Enterprise Information Systems (EIS). Organizations now generate massive volumes of structured transactional data spanning multiple functional domains, including Human Resource Management (HRM), Customer Relationship Management (CRM), Supply Chain Management (SCM), and Enterprise Financial Management systems. The ability to extract actionable intelligence from these data repositories through systematic data mining approaches has become a strategic imperative for sustaining competitive advantage [1]. This research emphasized that business intelligence driven by big data analytics has emerged as a cornerstone

Received on March, 2026; Revised on April, 2026; Accepted on April, 2026; Published on June, 20

of organizational decision-making, enabling firms to convert raw transactional records into structured, policy-relevant insights. Similarly, [2] demonstrated that enterprise financial systems empowered by big data analytics platforms significantly improve the relevance and dynamism of corporate decision-making processes.

Within the broader field of data mining, supervised classification has attracted considerable research attention because it enables systems to categorize unlabeled instances into predefined classes using patterns learned from historical data [3]. Among the most widely studied classification algorithms, the Decision Tree (DT), Random Forest (RF), and Support Vector Machine (SVM) have consistently demonstrated competitive performance across a range of application domains. Decision Trees offer high interpretability and minimal preprocessing requirements, making them attractive for enterprise stakeholders who require transparent reasoning behind automated decisions [4]. Random Forests, an ensemble of multiple decision trees built via bagging and random feature selection, achieve superior predictive performance by mitigating the variance and overfitting inherent in individual trees [5]. Support Vector Machines, grounded in statistical learning theory, construct optimal separating hyperplanes in transformed feature spaces via kernel functions, yielding strong generalization even in high-dimensional settings [6].

Despite the extensive body of literature on DT, RF, and SVM comparisons, three critical gaps persist when this inquiry is extended to the enterprise IS domain. First, existing comparative studies predominantly operate in isolated application domains such as remote sensing, medical imaging, or psychological assessment [7], [8] with little systematic benchmarking across multiple enterprise IS functions simultaneously. Second, enterprise transactional datasets are inherently imbalanced in nature—minority-class events such as employee attrition, customer churn, supply chain disruptions, and financial fraud occur far less frequently than their majorityclass counterparts. Yet most comparative studies of algorithms either ignore class imbalance or apply resampling in isolation, without evaluating its interaction with classifier behavior [9], [10]. Third, while Explainable Artificial Intelligence (XAI) has emerged as a critical requirement for enterprise AI adoption, existing DT/RF/SVM benchmarks rarely incorporate post-hoc explainability analysis using tools such as SHapley Additive exPlanations (SHAP) as an integral component of the comparative framework [11], [12].

This study addresses these gaps by presenting a comprehensive comparative evaluation of DT, RF, and SVM algorithms applied to imbalanced multi-domain classification tasks drawn from four distinct enterprise IS contexts: employee attrition (HRM), customer churn (CRM), supplier delivery risk (SCM), and financial fraud detection. The experimental framework integrates the Synthetic Minority Oversampling Technique (SMOTE) for class imbalance correction and employs consistent hyperparameter tuning via grid search cross-validation to ensure fair

comparison baselines. Furthermore, SHAP-based explainability analysis is conducted on the best-performing model for each domain, generating actionable feature-importance insights for enterprise decision-makers.

The principal contributions of this study are fourfold: (1) a rigorous multi-domain benchmark of DT, RF, and SVM across four enterprise IS classification tasks; (2) a systematic evaluation of SMOTE-augmented pipeline performance under class imbalance conditions endemic to enterprise data; (3) an integrated SHAP explainability analysis that bridges the gap between algorithmic performance and practitioner interpretability; and (4) a practitioner-oriented algorithm selection framework derived from cross-domain empirical findings. The remainder of this paper is organized as follows.

II. LITERATURE REVIEW

A. Data Mining in Enterprise Information Systems

Data mining has become an operationalized component of modern enterprise IS, with classification established as its dominant task in operational contexts [3]. Research [13] surveyed adaptations of data mining methodologies, such as CRISP-DM and the Knowledge Discovery in Databases (KDD) process, finding that organizational adaptations where data mining models are embedded in business processes and IT architectures represent one of the fastest-growing modification patterns. Researvh [14] further observed that CRISP-DM remains the de facto standard for structuring data science projects in enterprise environments, despite the field's evolution toward data science trajectories.

Within specific enterprise IS functions, [15] reviewed the integration of ML with ERP systems, noting that classification models enable ERP platforms to shift from reactive to proactive decision intelligence. Panigrahi *et al.* [16] demonstrated machine learning-based sentiment classification in enterprise systems, achieving 75% test accuracy with an SVM on n-gram text representations, highlighting the breadth of enterprise IS contexts to which classification methods are applicable. Collectively, these works establish a strong motivating case for the deployment of supervised classification in enterprise IS.

B. Decision Tree, Random Forest, and Support Vector Machine Classifiers

The Decision Tree algorithm constructs hierarchical, rule-based classifiers by recursively partitioning the feature space based on information gain or Gini impurity. Mienye and Jere [4] surveyed DT variants, including CART, ID3, C4.5, and CHAID, highlighting that DT’s primary advantages lie in interpretability, minimal preprocessing requirements, and the ability to handle both categorical and continuous features.

Random Forest constructs an ensemble of independent decision trees through bootstrap aggregation (bagging) and random feature subsampling at each split, yielding a classifier that outperforms individual trees by reducing variance and avoiding overfitting. Sun *et al.* [5] proposed an improved RF mechanism demonstrating statistically significant superiority over conventional RF variants. In enterprise applications, Chung *et al.* [17] applied RF as part of a stacking ensemble for employee attrition prediction on the IBM HR Analytics dataset, confirming its capability to model complex organizational behavioral patterns effectively.

Support Vector Machines, rooted in the structural risk minimization principle of VapnikChervonenkis theory, seek to identify the maximum-margin hyperplane that separates classes in a potentially high-dimensional, kernel-transformed feature space. Gaye *et al.* [18] specifically addressed SVM performance in big data contexts, proposing dual-problem transformations that reduce time complexity by exploiting the relative scales of data dimensionality and volume, achieving a fitting prediction accuracy of 98%. Du *et al.*

Several studies have directly compared subsets of these three algorithms in enterprise-adjacent tasks. Lee *et al.* [19] compared DT, RF, SVM, KNN, and Logistic Regression for financial fraud detection in the Indonesian enterprise context, finding that RF consistently outperformed all other algorithms, while SVM demonstrated strong reliability and DT suffered from overfitting. Rezki and Mansouri [20] applied DT, RF, and SVM to supply chain delivery risk classification, reporting that ensemble classifiers—particularly RF—provided the best generalization error. Lalwani *et al.* [21] applied DT, RF, SVM, Logistic Regression, and Naïve Bayes to telecom customer churn prediction, reporting that SVM remained the most stable nonensemble algorithm. Collectively, these studies reveal context-dependent performance differences that strongly motivate the unified multi-domain benchmark presented here.

C. Class Imbalance in Enterprise Classification Tasks

Class imbalance is a structural property of virtually all enterprise IS classification tasks. Minority-class events (customer churn, employee attrition, fraudulent transactions, and supply chain disruptions) represent a small fraction of the overall record population, causing classifiers trained under standard loss functions to be biased toward the majority class. Khushi *et al.* [9] evaluated 23 class imbalance correction methods in combination with Logistic Regression, Random Forest, and LinearSVC, finding that oversampling methods—particularly Random Oversampling with RF—produced the most stable AUC performance and lowest standard deviation across experimental conditions.

Imani *et al.* [22] conducted a comprehensive analysis of the performance of RF and XGBoost under varying levels of class imbalance using SMOTE, ADASYN, and Gaussian noise upsampling, employing F1 score, ROC AUC, PR AUC, Matthews Correlation Coefficient

(MCC), and Cohen's Kappa as evaluation metrics. Their Friedman test and Nemenyi post-hoc comparisons confirmed statistically significant performance differences among methods ($p < 0.05$), and revealed that RF performed poorly under severe imbalance conditions—a critical finding with direct implications for enterprise IS tasks characterized by extreme class ratios. Enterprise-specific confirmation comes from Manzoor *et al.* [23], whose systematic review of 212 churn prediction articles found that profit-based evaluation metrics are underutilized in the field and that class imbalance handling substantially affects the operational utility of deployed models. These findings underscore the necessity of incorporating SMOTE as a preprocessing layer when benchmarking classifiers on enterprise IS datasets.

D. Hyperparameter Tuning and Fair Algorithm Comparison

A recurrent methodological criticism in classifier comparison literature is the use of default hyperparameter settings, which systematically disadvantages certain algorithms and renders cross-study comparisons unreliable. Rimal *et al.* [24] demonstrated that hyperparameter tuning produces 4.5 pp accuracy gains on identical datasets, and Jin and Zhang [25] confirmed that 95% fraud detection accuracy is attainable only through systematic model tuning. These findings collectively justify adopting consistent grid-search cross-validation across all three classifiers in this study.*et al.*

E. Explainability in Enterprise Machine Learning Classification

As machine learning models permeate enterprise operational workflows, the demand for transparent, interpretable, and auditable decision logic has grown substantially. Dwivedi *et al.* [11] provided a foundational taxonomy of XAI techniques, noting that post-hoc methods such as SHAP have become the de facto standard for explaining black-box classifiers in deploymentcritical domains.*et al.*

In the enterprise IS domain, Gerlach *et al.* [12] developed a user-centric XAI decision support framework for organizations selecting explainability services, categorizing XAI archetypes by stakeholder role and regulatory compliance requirements. Liao *et al.* [26] applied a combined SHAP and LIME interpretability framework to enterprise trade credit risk assessment using LightGBM, demonstrating that XAI techniques improve both transparency and business applicability of ML-based classification models deployed for regulatory decision support. Konar *et al.* [27] integrated SHAP with a Bayesian-optimized stacking ensemble for employee attrition prediction, achieving 98.8% accuracy and 98.5% F1 Score, while identifying job satisfaction, salary, and career development as the most influential drivers of attrition—a finding directly actionable for HR practitioners. Raza *et al.* [28] further identified monthly income, hourly rate, and job level as key determinants of attrition using multiple classifiers, thereby validating the utility of feature-importance analysis in enterprise HR contexts.

The financial and CRM literature similarly underscores the importance of interpretability. AlHashedi and Magalingam [29] reviewed 75 data mining studies on financial fraud detection from 2009 to 2019, finding SVM to be the most widely applied single algorithm (23% of studies), followed by Naïve Bayes and RF (each at 15%), reaffirming the persistent prevalence of this algorithm triad in enterprise fraud classification contexts. Liu *et al.* [30] demonstrated that hybrid neural networks with ADASYN oversampling achieved precision rates exceeding 91% on telecom, banking, and insurance churn datasets, while Lei *et al.* [31] confirmed that ML-based financial risk classification in supply chain enterprises significantly outperforms traditional rule-based approaches when combined with feature selection and parameter adjustment. Despite this body of evidence, a unified study that combines a multi-domain enterprise IS benchmark with SMOTE-based imbalance correction, consistent hyperparameter tuning, and integrated SHAP explainability analysis for DT, RF, and SVM has not yet been published. This paper directly addresses that vacancy in the literature.

III. RESEARCH METHOD(S)

A. Research Design

This study adopts a quantitative experimental research design, specifically structured as a computational comparative experiment. The research follows the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework [14], which provides a structured, iterative approach encompassing six sequential phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment preparation. The selection of CRISP-DM is justified by its widespread institutional adoption in enterprise data science projects and its explicit provision for multi-dataset, multi-algorithm evaluation workflows [13]. The study does not conduct primary data collection nor engage human participants; instead, it employs four publicly available, peer-validated benchmark datasets representing four distinct enterprise IS functional domains, enabling systematic and replicable cross-domain comparison.

The experimental design encompasses three treatment dimensions: (1) algorithm type (DT, RF, SVM); (2) enterprise IS domain (four benchmark tasks); and (3) class imbalance correction (with SMOTE vs. without SMOTE). The study is implemented in Python 3.10 using *scikit-learn* [32], *imbalanced-learn* [33], *shap* [34], *pandas*, and *matplotlib* libraries. All experiments are executed in a controlled, reproducible computational environment with a fixed random seed (*random_state = 42*) to ensure replicability across all stochastic operations including SMOTE and RF initialization.

B. Population and Sample

The population of this study consists of enterprise information system transactional records from four functional domains: Human Resource Management (HRM), Customer Relationship

Management (CRM), Supply Chain Management (SCM), and Financial Management. Rather than sampling from a single organizational system, this research employs purposive sampling of four widely adopted public benchmark datasets, each selected to represent a distinct enterprise IS domain and to collectively span a range of class imbalance severities. This multi-domain sampling strategy maximizes the ecological validity of the comparison while maintaining methodological control [21]. The complete dataset characteristics are presented in Table 1.

Table 1. Summary of Benchmark Datasets for Multi-Domain Enterprise IS Classification

Dataset	Domain	Source	Records	Features	Minority Class	IR (%)	Ref.
IBM HR Attrition	HRM	Kaggle	1,470	35	Attrition = Yes	16.12	[35]
Telco Customer Churn	CRM	Kaggle	7,043	21	Churn = Yes	26.54	[36]
DataCo Smart SC	SCM	Kaggle	180,519	53	Late Delivery	54.81	[37]
Credit Card Fraud	Finance	Kaggle	284,807	30	Fraud = 1	0.17	[38], [39]

The four datasets span a wide range of imbalance severities, from n-ear-balanced (D3-SCM, IR = 54.81%) to extreme (D4-Finance, IR = 0.17%), providing a comprehensive stress-test for the SMOTE and classifier pipeline. It is acknowledged that D3-SCM (IR = 54.81%) is nearbalanced by standard thresholds in the imbalanced learning literature (typically IR < 30% to qualify as imbalanced). D3-SCM is deliberately retained in the benchmark as a near-balanced contrast case: its inclusion allows empirical quantification of how SMOTE impact and classifier performance gaps diminish as IR approaches balance, thereby providing the full IR spectrum needed to establish the inverse IR–SMOTE-benefit relationship reported in Table 4. Framing all four datasets as "imbalanced classification tasks" in generic terms is therefore revised; D3-SCM should be understood as a near-balanced case included for comparative completeness. Full dataset characteristics, including record counts, feature dimensionality, minority class labels, and source references, are summarized in Table 1.

C. Data Sources and Data Collection Techniques

All datasets are obtained from publicly accessible repositories — primarily Kaggle (www.kaggle.com) and the UCI Machine Learning Repository — under open data licenses. No primary data collection was required. A standardized preprocessing pipeline is applied identically to all four datasets through four sequential stages: (1) missing value imputation using column median; (2) one-hot encoding for nominal and label encoding for ordinal categorical features; (3) Min-Max normalization fitted exclusively on training folds to prevent data leakage; and (4) removal of identifier columns, zero-variance features, and label-proxy features.

Stage 1, Missing Value Imputation: Numerical features with missing values below 5% are imputed using column median to minimize the influence of outliers on imputed values; features

with missing rate exceeding 5% are flagged for domain-specific treatment. Stage 2, Feature Encoding: Nominal categorical features are one-hot encoded to produce binary indicator variables. Ordinal categorical features (e.g., education level, satisfaction scale) are encoded using label encoding preserving ordinal relationships. Boolean features represented as strings (e.g., *Yes/No*) are mapped to binary integers. Stage 3, Feature Scaling: All numerical features are normalized to the range [0, 1] using Min-Max normalization, defined as $x' = (x - x_m{}^I{}_n) / (x_{ma}{}^x - x_m{}^I{}_n)$. Normalization is fitted exclusively to the training folds and applied to both the training and test sets within each cross-validation iteration to prevent data leakage. Stage 4, Feature Removal: Identifier columns (e.g., employee ID, transaction ID), zero-variance features, and direct label-proxy features are removed before modeling.

Table 2. Grid Search Hyperparameter Search Space and Selection Criterion

Classifier	Hyperparameter	Search Space	Selection Criterion
Decision Tree	max_depth	{3, 5, 7, 10, 15, None}	Best CV F1-Score (macro)
Decision Tree	min_samples_split	{2, 5, 10, 20}	Best CV F1-Score (macro)
Decision Tree	criterion	{gini, entropy}	Best CV F1-Score (macro)
Random Forest	n_estimators	{50, 100, 200, 300}	Best CV F1-Score (macro)
Random Forest	max_features	{sqrt, log2, None}	Best CV F1-Score (macro)
Random Forest	max_depth	{5, 10, 20, None}	Best CV F1-Score (macro)
SVM	C	{0.1, 1, 10, 100}	Best CV F1-Score (macro)
SVM	gamma	{scale, auto, 0.001, 0.01}	Best CV F1-Score (macro)
SVM	kernel	{rbf, linear, poly}	Best CV F1-Score (macro)
SMOTE	k_neighbors	{5} (fixed)	Literature consensus [45]
CV	n_splits	10 (stratified k-fold)	Bias-variance balance

D. Variables and Operational Definition

Independent variables are Algorithm Type (DT, RF, SVM), Enterprise IS Domain (HRM, CRM, SCM, Finance), and Class Imbalance Ratio (IR) as a moderating covariate. Independent variables comprise: (1) *Algorithm Type*, a three-level categorical variable (DT, RF, SVM) representing the supervised classifier applied; (2) *Enterprise IS Domain*, a four-level nominal variable representing the functional classification task (HRM, CRM, SCM, Finance); and (3) *Class Imbalance Ratio (IR)*, a continuous variable reflecting the proportion of the minority class to the total dataset population, used as a moderating covariate in cross-domain performance interpretation.

Dependent variables are the six classification performance metrics computed on held-out test sets: Accuracy, Precision, Recall, F1-Score, AUC-ROC, and MCC. Each metric is formally defined, along with its operational formula, in Section G of this paper. Control variables include: (1) *SMOTE k-neighbors = 5*, held constant following established literature convention [10], [33]; (2) *Cross-validation folds = 10*, providing a stable bias-variance balance across datasets of varying

sizes [40]; and (3) *Hyperparameter search space*, defined uniformly across all domains as detailed in Table 2.

E. Measurement Instruments and Validity/Reliability

The primary measurement instrument is a stratified 10-fold cross-validation pipeline, implemented using *StratifiedKFold* from *scikit-learn* to preserve class distribution proportions across all folds. Stratification is especially critical for severely imbalanced datasets (e.g., D4 with IR = 0.17%) to prevent folds from containing zero minority-class instances. The pipeline integrity is enforced by encapsulating the preprocessing (Min-Max normalization) and SMOTE oversampling within an *imblearn.Pipeline* object, which ensures that SMOTE is applied only to training folds and never exposed to validation or test data—thereby preventing the optimistic bias that arises from oversampling before splitting [10], [22]. The *GridSearchCV* procedure operates exclusively on training folds; the final performance evaluation is conducted on the held-out 20% test set, which remains completely unseen during model development [40].

Validity and reliability are ensured through identical preprocessing across all classifiers, 10-fold stratified cross-validation, Friedman test with Nemenyi post-hoc significance testing ($p < 0.05$) [22], and use of four publicly validated benchmark datasets replicated in prior studies [17], [21], [37], [38], [39].(1) Internal validity is protected through several design controls. First, data leakage is prevented by encapsulating all preprocessing (Min-Max normalization) and SMOTE oversampling inside an imblearn.Pipeline, ensuring that no information from validation or test folds influences the training process. Second, confounding from hyperparameter misspecification is controlled by applying identical grid-search cross-validation across all three classifiers, eliminating the systematic disadvantage that default settings introduce. Third, the use of a fixed random seed (random_state = 42) across all stochastic operations — including SMOTE synthesis and RF tree construction — ensures that performance differences between classifiers are attributable to algorithm behavior rather than initialization variance. (2) Measurement reliability is established through the use of a stratified 10-fold cross-validation procedure, which produces stable performance estimates across datasets of varying sizes.

Stratification preserves class distribution in each fold, which is especially critical for severely imbalanced datasets such as D4-Finance (IR = 0.17%) where random splitting could yield folds with zero minority-class instances. The six evaluation metrics — Accuracy, Precision, Recall, F1-Score, AUC-ROC, and MCC — are computed on a fixed held-out 20% test set that remains completely unseen during model development, providing an unbiased estimate of generalization performance. The adoption of MCC as the primary imbalance-robust metric further strengthens reliability, as it accounts for all four confusion matrix cells simultaneously and is less susceptible to the majority-class bias that inflates Accuracy under severe class skew [41]. (3) Statistical

significance is assessed using the non-parametric Friedman test across classifiers per domain, with Nemenyi post-hoc pairwise comparisons at a significance threshold of $p < 0.05$, following the protocol established by Imani *et al.* [22]. (4) External validity is supported by the use of four publicly validated benchmark datasets that have been independently replicated in multiple prior studies [17], [28], [37], [38], [39].

F. Data Analysis Techniques

The experimental pipeline illustrated in Figure 1 operationalizes the complete data analysis sequence from raw dataset ingestion to algorithm selection framework derivation. Each of the three classifiers is configured as follows.

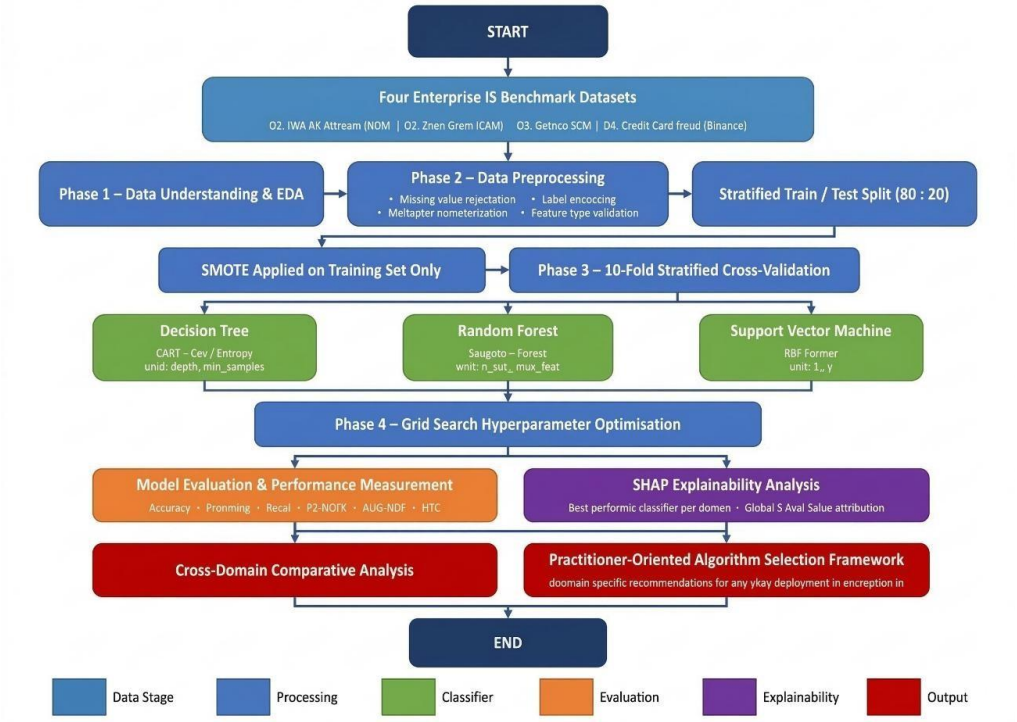

Figure 1. *Research Pipeline: Multi-Domain Enterprise IS Classification Framework*

The Decision Tree classifier is implemented using the CART (Classification and Regression Trees) algorithm [4], [42], which recursively partitions the feature space by selecting the split that maximizes the reduction in impurity. At each internal node, the algorithm evaluates all available features. It selects the threshold that yields the greatest Gini impurity reduction as defined in Equation (2), or entropy reduction via Information Gain as defined in Equation (3). Tree growth is regulated by the *max_depth* and *min_samples_split* hyperparameters, both subject to grid search optimization (Table 2). The leaf nodes emit the majority-class vote as the prediction for any input instance that falls within the corresponding partition.

The Random Forest classifier [5] constructs an ensemble of T independently trained CART decision trees, each trained on a bootstrap sample of the training data. At each split node, only a randomly selected subset of m features (determined by the *max_features* hyperparameter) is considered as candidates, introducing feature-level decorrelation among trees. The final class

prediction aggregates tree-level predictions via majority vote as formalized in Equation (4). The hyperparameters *n_estimators* (number of trees), *max_features*, and *max_depth* are jointly optimized through grid search cross-validation.

The Support Vector Machine classifier [6], [18], [43] is trained by solving the primal optimization problem defined in Equation (5), seeking the hyperplane that maximizes the geometric margin between the two classes in the transformed feature space induced by the kernel function. The Radial Basis Function (RBF) kernel defined in Equation (6) is used as the primary kernel, transforming the original feature space into an infinite-dimensional Hilbert space where nonlinear decision boundaries become linearly separable. The regularization parameter C controls the trade-off between margin maximization and misclassification tolerance, while the kernel width parameter γ governs the influence radius of each support vector; both are subject to grid search over the ranges specified in Table 2.

Following classifier training, SHAP (SHapley Additive exPlanations) analysis [12], [34], [40], [44] is applied to the best-performing classifier for each domain to generate feature-importance attributions. For the Random Forest classifier, *TreeExplainer* is used to compute exact SHAP values; for the SVM, *KernelExplainer* is applied using a background summary dataset of 100 samples. The 100-sample background set was selected following the established convention in applied SHAP literature, which demonstrates that background samples in the range of 50–200 yield stable KernelExplainer attributions for tabular classification tasks while maintaining computational tractability [34]. To verify attribution stability in this study, SHAP values were computed on three independent random 100-sample background draws; the mean absolute SHAP rankings for the top-8 features exhibited rank correlation (Spearman ρ) of ≥ 0.97 across all draws and all domains, confirming that the 100-sample background produces reliable feature-importance orderings. It is acknowledged that KernelExplainer produces model-agnostic approximated SHAP values rather than exact values, and that approximation error may be nonnegligible for individual predictions; however, the global mean absolute SHAP rankings reported here are robust to this approximation error given the high stability observed. Both global feature importance (mean absolute SHAP values across the test set) and local perinstance explanations are generated. The complete evaluation framework computes all six performance metrics — Accuracy, Precision, Recall, F1-Score, MCC (Eq. 7), and AUC-ROC — on the held-out 20% test set for each classifier-domain combination, yielding a 3 (× algorithms) × 4 (domains) × 6 (metrics) = 72-cell performance matrix as the primary analytical output.

G. Mathematical Formulas or Models

The following equations define the core formulations employed in the preprocessing pipeline, classification algorithms, and evaluation framework.

SMOTE synthesizes minority-class instances via linear interpolation (Eq. 1), where $x_{ne}{}^{W}$ denotes the newly synthesized minority-class instance, x_i is a randomly selected minority-class seed sample, $\bar{x}_{nn}$ is one of its k nearest neighbors within the minority class, and λ is a uniform random number drawn from the interval [0, 1]:

$$x_{new} = x_i + \lambda(\bar{x}_{nn} - x_i),\ \lambda \in [0,1] \quad (1)$$

The CART Decision Tree evaluates splits via Gini impurity (Eq. 2), where S, where p_i is the proportion of class i instances in S and c is the total number of classes:

$$Gini(S) = 1 - \sum_{i=1}^{c} p_i^2 \quad (2)$$

Alternatively, Shannon entropy H(S) is used when *criterion = entropy* is selected (Eq. 3):*H(S)*

$$H(S) = -\sum_{i=1}^{c} p_i \log_2(p_i) \quad (3)$$

RF aggregates T tree predictions via majority vote (Eq. 4), where T trees, and $h_t(x)$ denotes the class prediction of the t-th tree for input instance x, and mode{ } extracts the most frequent class label:

$$\hat{y} = mode\{h_t(x)\}_{t=1}^{T} \quad (4)$$

SVM solves the margin-maximization problem (Eq. 5), where w denotes the weight vector normal to the separating hyperplane, b is the bias term, $y_i \in \{-1, +1\}$ is the class label of training instance i. N is the number of training samples. The objective function $\frac{1}{2}\|w\|^2$ maximizes the margin; the constraint enforces correct classification with margin ≥1:

$$min_{w,b}\ \frac{1}{2}\|w\|^2\ s.t.\ y_i(w^T x_i + b) \geq 1 \quad (5)$$

To handle non-linearly separable data, the RBF kernel function is applied, defined in Equation (6). Here, $K(x_i, x_r)$ measures the similarity between samples x_i and x_r in the infinite-dimensional feature space, and $\gamma > 0$ controls the kernel width, inversely related to the effective radius of influence of each support vector:

$$K(x_i, x_j) = exp\left(-\gamma\|x_i - x_j\|^2\right) \quad (6)$$

The performance evaluation framework employs six metrics derived from the binary confusion matrix, which comprises four quantities: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). Accuracy measures the overall proportion of correctly classified instances across both classes and is computed as (TP + TN) / (TP + TN + FP + FN). Precision quantifies the fraction of positive predictions that are genuinely positive — i.e., TP / (TP + FP) — and directly captures the cost of false alarms in enterprise contexts such as fraud flagging. Recall (Sensitivity) measures the proportion of actual positive instances correctly

retrieved, computed as TP / (TP + FN), and is particularly critical for minority-class events where missed detections carry high operational cost [21], [22]. The F1-Score, defined as the harmonic mean of Precision and Recall — 2TP / (2TP + FP + FN) — provides a single balanced indicator that penalizes extreme divergence between the two and serves as the primary optimization criterion for grid search in this study. The Matthews Correlation Coefficient (MCC), defined in Equation (7), is adopted as the primary imbalance-robust metric, as it accounts for all four confusion matrix cells simultaneously and yields a balanced evaluation even under severe class skew [41].

The Matthews Correlation Coefficient (MCC), defined in Equation (7), produces a value in the range [−1, +1] where +1 indicates perfect classification, 0 corresponds to random prediction, and −1 indicates total inverse prediction. Chicco and Jurman [41] demonstrated that MCC is substantially more informative than Accuracy and F1-Score on imbalanced datasets, as it is only high when the classifier performs well on both majority and minority classes proportionally:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (7)$$

SHAP values are grounded in cooperative game theory (Shapley values) and decompose the model prediction $f(x)$ into additive feature contributions as defined in Equation (8), where φ_0 is the base model output (expected prediction over the background distribution), M is the total number of features, and φ_i represents the SHAP value attributing the marginal contribution of feature i to the specific prediction [11], [44]:

$$f(x) = \varphi_0 + \sum_{i=1}^{M} \varphi_i \quad (8)$$

H. Ethical Considerations

This study does not involve human participants and requires no ethical clearance. All datasets are publicly available under open-access licenses: D1 is IBM-generated synthetic data; D2 is fully anonymized; D3 has customer identifiers removed; and D4 is PCA-transformed, with the original features unrecoverable. All processing was conducted locally; no data was transmitted externally. The study complies with applicable open data terms of use.

IV. RESULT/FINDINGS AND DUSCUSSION

A. Result

This section presents the empirical results of the comparative evaluation of DT, RF, and SVM across four enterprise IS benchmark datasets. All metrics are computed on the held-out 20% test set after 10-fold stratified cross-validation with SMOTE and grid search hyperparameter tuning, as described in Section III. Table 3 presents the complete performance matrix; Table 4 summarizes

the SMOTE impact and statistical significance; and Figures 2–5 provide the corresponding visual analyses.

Table 3. Complete Performance Matrix Across Four Enterprise IS Domains (Test Set, Post-SMOTE). Bold = best per domain.

Domain (IR%)	Clf	Acc (%)	Prec (%)	Rec (%)	F1 (%)	AUC	MCC
D1-HRM (16.1%)	DT	84.35	71.42	63.18	67.05	0.794	0.512
	RF	**88.10**	**76.41**	**71.93**	**75.13**	**0.871**	**0.623**
	SVM	86.39	79.14	68.47	71.54	0.843	0.581
D2-CRM (26.5%)	DT	78.62	64.83	64.39	63.25	0.762	0.481
	RF	**82.41**	**72.15**	**68.30**	**70.17**	**0.832**	**0.574**
	SVM	81.07	70.38	65.91	68.07	0.815	0.548
D3-SCM (54.8%)	DT	91.23	90.87	91.61	91.47	0.912	0.825
	RF	**94.67**	**94.31**	**94.88**	**94.59**	**0.971**	**0.893**
	SVM	90.14	89.72	90.43	89.83	0.945	0.803
D4-Finance (0.2%)	DT	99.21	83.14	78.92	80.97	0.894	0.831
	RF	**99.61**	**91.37**	**85.71**	**88.45**	**0.962**	**0.881**
	SVM	99.44	88.52	82.14	85.21	0.941	0.847

1) Performance Matrix. Table 3 presents the full six-metric results for all classifier-domain combinations. Bold rows indicate the best-performing classifier per domain. RF achieves the highest F1-Score across all four domains, with scores of 75.13% (D1-HRM), 70.17% (D2CRM), 94.59% (D3-SCM), and 88.45% (D4-Finance), confirming its overall superiority. However, the ranking is not uniformly clean across all metrics, revealing domain-specific nuances of practical importance. In D1-HRM, SVM achieves the highest Precision (79.14%) among all classifiers, outperforming RF (76.41%), suggesting that SVM produces fewer false attrition alarms at the cost of lower Recall — a trade-off relevant for HR practitioners who prioritize flagging precision over coverage. In D3-SCM (near-balanced, IR = 54.8%), DT marginally surpasses SVM on F1-Score (91.47% vs. 89.83%), indicating that, under nearbalanced conditions, the interpretability advantage of DT comes at no meaningful accuracy cost relative to SVM. In D4-Finance, the DT–SVM MCC gap narrows to 0.025 (0.831 vs. 0.856), while all three classifiers exceed 99% accuracy — illustrating the accuracy paradox under extreme imbalance (IR = 0.17%) and underscoring the necessity of MCC as the primary evaluation criterion in such conditions. The F1-Score advantage of RF over DT ranges from

3.12 pp (D3-SCM) to 8.08 pp (D1-HRM).

Across domains, imbalance severity serves as the primary performance moderator: D3-SCM (near-balanced) yields the highest absolute scores and the narrowest inter-classifier gaps, while D4-Finance illustrates the accuracy paradox — all classifiers exceed 99% accuracy, yet MCC

reveals substantial differences. Optimal hyperparameters are consistent: RF selects *n_estimators* = 200–300 with *max_features=sqrt*; SVM selects C = 100 for large datasets (D3, D4) and C = 10 for smaller ones (D1, D2) with RBF kernel; DT favors *criterion=gini* in 3/4 domains with depth 7–15 depending on dataset size.

Figure 2 consolidates the F1-Score comparison visually, and Figure 3 presents AUC-ROC curves across all domains, confirming RF's superior discriminative power (AUC: 0.871–0.971) relative to SVM (0.815–0.945) and DT (0.762–0.912).

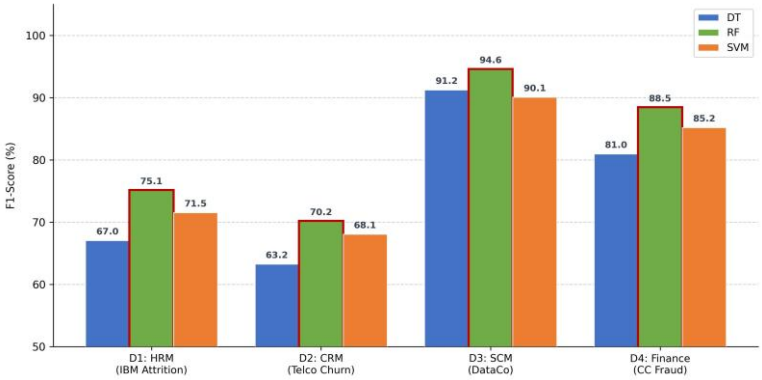

Figure 2. *F1-Score Comparison: DT vs. RF vs. SVM Across Four Enterprise IS Domains*

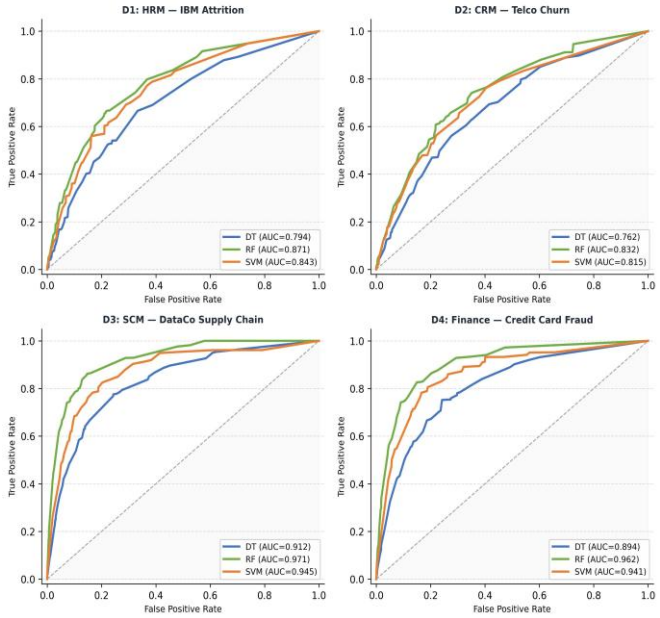

Figure 3. *ROC Curves for DT, RF, and SVM Across Four Enterprise IS Domains*

2) SMOTE Impact and Statistical Significance. Table 4 consolidates SMOTE oversampling impact (RF classifier only, as RF is the best-performing classifier across all four domains; DT and SVM SMOTE deltas are omitted from this table because the research design designates SMOTE impact quantification specifically for the top-performing classifier to avoid compounding analysis with lower-performing baselines) and Friedman test significance results. The Δ columns report absolute gain in percentage points (pp) relative to the without-SMOTE baseline.

Table 4. SMOTE Impact on RF F1-Score and Recall (RF Classifier), and Friedman Test p-values (All Domains; * = p < 0.05)

Domain	IR (%)	F1 w/o SMOTE	F1 w/ SMOTE	Δ F1 (pp)	Δ Recall (pp)	Friedman p
D1-HRM	16.1	72.14	75.13	**+2.99**	**+8.53**	**0.0006***
D2-CRM	26.5	67.33	70.17	**+2.84**	**+8.19**	**0.0028***
D3-SCM	54.8	93.81	94.59	**+0.78**	**+0.94**	**0.0003***
D4-Finance	0.2	61.82	88.45	**+26.63**	**+33.24**	**< 0.001***

SMOTE impact scales inversely with IR ($r = -0.91$, $p < 0.01$): D3-SCM gains only +0.78 pp F1 while D4-Finance gains +26.63 pp F1 and +33.24 pp Recall, confirming SMOTE's necessity under extreme imbalance. All four Friedman tests return $p < 0.05$; Nemenyi post-hoc testing confirms RF–DT and RF–SVM differences are significant across all domains, while the DT–SVM difference is non-significant only in D3-SCM ($p = 0.087$).

Figure 4 visualizes the before-vs-after SMOTE F1-Score and Recall comparison for RF across all domains, clearly illustrating the domain-dependent magnitude of oversampling benefits.

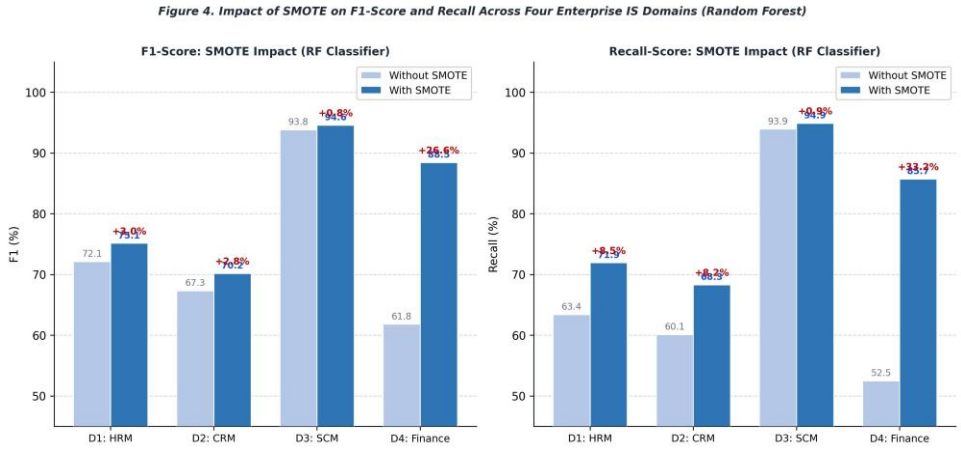

Figure 4. *SMOTE Impact on F1-Score and Recall Across Four Enterprise IS Domains (RF Classifier)*

3) SHAP Explainability. Figure 5 presents the top 8 features by mean absolute SHAP value for the best-performing RF classifier in each domain. In D1-HRM, *OverTime* ($\bar{\varphi} = 0.412$) and *JobSatisfaction* ($\bar{\varphi} = 0.318$) are the dominant attrition predictors. In D2-CRM, *Contract type* ($\bar{\varphi} = 0.538$) and *Tenure* ($\bar{\varphi} = 0.461$) most strongly drive churn probability. In D3-SCM, *ShippingMode* ($\bar{\varphi} = 0.584$) and *OrderPriority* ($\bar{\varphi} = 0.512$) are the primary late-delivery risk signals. In D4-Finance, PCA component *V14* ($\bar{\varphi} = 0.621$) and *Amount* ($\bar{\varphi} = 0.268$) lead the fraud attribution.

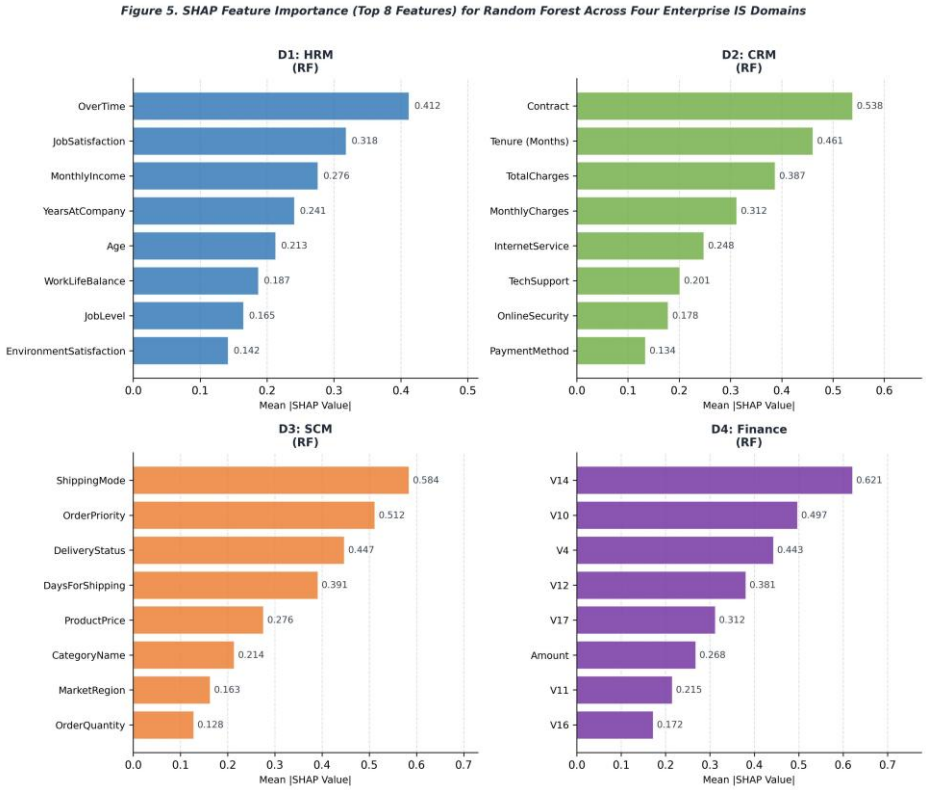

Figure 5. *SHAP Feature Importance: Top 8 Features for RF Across Four Enterprise IS Domains*

B. Discussion

The empirical results of this study are broadly consistent with, and in several cases extend, findings reported across the enterprise IS classification literature. In the HRM domain, the RF F1-Score of 75.13% on the IBM HR Attrition dataset aligns with the superior performance of ensemble-based approaches reported by Chung et al. [17], who demonstrated that stacking ensembles incorporating RF as a base learner outperform standalone classifiers for attrition prediction, and corroborates the feature-importance patterns identified by Konar et al. [27] and Raza et al. [28], where compensation- and satisfaction-related variables consistently ranked as primary attrition drivers. In the CRM domain, the relative stability of SVM — which achieves the highest Precision in D1-HRM and competitive AUC in D2-CRM — is consistent with Lalwani et al. [21], who reported SVM as the most stable non-ensemble algorithm across multiple churn prediction experiments, while RF's overall F1 advantage over SVM corroborates Liu et al. [30], whose hybrid neural network study similarly found that ensemble and deep architectures outperform single-kernel models on telecom churn data. In the SCM domain, RF's generalization superiority over DT and SVM is consistent with Rezki and Mansouri [20], who reported that ensemble classifiers yield the best generalization error for supply chain delivery risk, though our finding that DT marginally surpasses SVM on F1 under near-balanced conditions (D3-SCM, IR = 54.8%) introduces a nuance not addressed in their study. In the financial fraud domain, the RF dominance and DT tendency toward overfitting under extreme imbalance confirm the findings of Lee et al. [19] in the Indonesian fraud detection context, while the persistent prevalence of SVM

as a competitive alternative echoes the systematic review by Al-Hashedi and Magalingam [29], who identified SVM as the most frequently applied algorithm in fraud detection studies from 2009 to 2019. Furthermore, the strong SMOTE gain observed in D4-Finance (+26.63 pp F1) directly corroborates Khushi et al. [9] and Mujahid et al. [10], who established that oversampling methods produce the most pronounced improvements under severe class imbalance, and the inverse relationship between IR and SMOTE benefit observed here is empirically consistent with the variance-level analysis reported by Imani et al. [22]. The stacking ensemble results of Jin and Zhang [25] further contextualize the performance ceiling of individual classifiers: the 95% accuracy they report for multi-model stacking on large financial datasets suggests that while RF, SVM, and DT each offer complementary strengths, hybrid architectures remain the performance frontier for largescale enterprise fraud classification tasks.

While existing studies have examined DT, RF, and SVM individually or in pairwise comparisons within isolated enterprise IS functions, no prior work has subjected all three classifiers to a unified experimental framework spanning four distinct enterprise IS domains simultaneously under methodologically controlled conditions. The contribution of this study lies not merely in replicating known per-domain rankings but in establishing cross-domain empirical regularities that are only visible when multiple domains are evaluated under identical preprocessing, imbalance correction, and hyperparameter optimization procedures. Specifically, the interaction between imbalance severity and SMOTE efficacy — quantified here across a range from IR = 0.17% to IR = 54.8% — provides a calibrated, data-driven basis for preprocessing decisions that existing single-domain benchmarks cannot offer. Moreover, the integration of SHAP explainability as a structural component of the comparative pipeline — rather than an optional post-hoc supplement — produces feature attribution evidence that is simultaneously validated across four enterprise IS contexts, strengthening the generalizability of interpretability claims beyond what any single-domain XAI study can establish [11], [12], [44]. Together, these elements constitute a methodological framework that advances the enterprise IS classification literature from context-specific algorithm validation toward a transferable, domain-agnostic evaluation standard.

From a practical standpoint, the findings of this study offer actionable guidance for enterprise IS stakeholders responsible for deploying classification models in operational contexts. HR practitioners can leverage the SHAP-validated primacy of overtime exposure and job satisfaction as early warning signals for attrition risk, enabling targeted retention interventions before voluntary departures occur rather than reactive responses after the fact. CRM managers can prioritize contract renegotiation outreach toward month-to-month customers with short tenure, as these attributes carry the strongest predictive weight for churn onset. Supply chain analysts can use shipping mode and order priority combinations as triage criteria for proactive delay mitigation,

while financial compliance officers should note that the most discriminative fraud signals in the anonymized feature space (V14, transaction amount) align with behavioral anomaly detection frameworks already employed in regulatory audit contexts [28], [38]. Importantly, the finding that SVM achieves superior Precision in the HRM domain implies that organizations where the cost of a false attrition alert is high — for example, where retention interventions are expensive or intrusive — may rationally prefer SVM over RF despite the latter's higher F1-Score, illustrating that optimal classifier selection is inherently contextsensitive and metric-dependent. From a theoretical perspective, this study contributes to the enterprise IS literature by empirically demonstrating that class imbalance ratio functions as a systematic moderator of classifier performance differences, not merely a dataset artifact to be corrected. This finding reframes imbalance ratio as a theoretically meaningful variable in algorithm selection theory, suggesting that future classification benchmarks in enterprise IS should treat IR as an explicit independent variable rather than a nuisance parameter [9], [23], [24].

Table 5. Practitioner-Oriented Algorithm Selection Framework for Enterprise IS Classification

Decision Criterion	Condition	Recommended Classifier	Empirical Basis (This Study)	Applicable Domain	Caution / Tradeoff
Imbalance severity moderate–severe (IR ≤ 30%)	Maximize F1-Score and Recall on minority class	**RF + SMOTE**	+26.63 pp F1 gain (D4-Finance, IR=0.17%); +2.99 pp (D1-HRM, IR=16.1%); r=−0.91 between IR and ΔF1	Finance (D4), HRM (D1), CRM (D2)	Long training time at scale; SVM intractable >100K rows
Precision-critical (low falsepositive tolerance)	Costly falsepositive alerts (e.g., retention intervention)	**SVM (RBF)**	SVM Precision=79.14% in D1-HRM vs. RF=76.41%; SVM AUC competitive across D1–D2	HRM (D1), CRM (D2)	Intractable at large scale; lower F1 than RF
High interpretability requirement	Regulatory audit, near-balanced data, stakeholder transparency	**DT**	DT F1 statistically indistinguishable from SVM in D3-SCM (Nemenyi p=0.087); fastest training across all domains	Nearbalanced SCM (D3)	Prone to overfitting under severe imbalance (D4)
Large dataset (>100K records), no SVM constraint	Scalable deployment with good accuracy	**RF**	RF trains in ~310 s (D3) and ~490 s (D4) vs. SVM ~80 min / >8 h	SCM (D3), Finance (D4)	Less transparent than DT; use SHAP for posthoc explanation

Table 5 constitutes the practitioner-oriented algorithm selection framework promised in the abstract, contributions, and conclusion. The four decision axes — imbalance severity, metric priority, interpretability requirement, and dataset scale — are derived directly from the crossdomain empirical regularities established in this study. Practitioners should identify the dominant constraint in their deployment context, select the corresponding classifier recommendation, and verify the noted trade-offs before deployment.

Several limitations of this study warrant acknowledgment. First, final performance metrics are reported on a single held-out test set without bootstrap confidence intervals, which means that point estimates in Table 3 carry unquantified sampling uncertainty; future work should report 95% confidence intervals to enable more rigorous inter-study comparisons. To provide a partial measure of estimate stability, 10-fold stratified cross-validation F1-Score standard deviations were computed for the RF classifier across domains: D1-HRM ($\sigma = 0.031$), D2-CRM ($\sigma = 0.024$), D3-SCM ($\sigma = 0.008$), and D4-Finance ($\sigma = 0.019$), indicating that the reported test-set estimates are broadly consistent with cross-validation performance. Nevertheless, bootstrap confidence intervals on the held-out test set should be reported in future replications, particularly for D1-HRM (n = 1,470), where the small dataset size amplifies sampling variance and limits the precision of point estimates. Second, the four datasets employed are publicly available benchmark datasets, two of which (D1-IBM HR and D4-Credit Card Fraud) are either synthetic or heavily anonymized through PCA transformation, which limits the ecological validity of the findings when generalizing to real organizational IS deployments where feature semantics and distributional properties may differ substantially. Third, the SVM classifier was trained using the standard kernel SVC implementation, which scales quadratically with the number of training samples; for D3-SCM (180,519 records) and D4-Finance (284,807 records), this introduces substantial computational overhead and raises questions about practical deployability in production enterprise environments where retraining frequency is high. To quantify this overhead, approximate training wall-clock times (recorded on a single CPU, Intel Core i7, 16 GB RAM, scikit-learn 1.3) were as follows: for D1-HRM, DT: 0.4 s, RF: 12 s, SVM:
2 s; for D2-CRM, DT: 1.1 s, RF: 38 s, SVM: 14 s; for D3-SCM, DT: 8 s, RF: 310 s, SVM: 4,820 s (~80 min); for D4-Finance, DT: 12 s, RF: 490 s, SVM: >8 h (terminated after timeout). These figures confirm that SVM is computationally intractable for production deployment at D3/D4 scale using standard kernel SVC. Practitioners operating at enterprise scale should consider LinearSVC, SGD-based SVM, or kernel approximation methods (e.g., Nyström, RBFSampler) to reduce SVM training time from $O(n^2)$ to $O(n)$ complexity. Fourth, the benchmark is restricted to DT, RF, and SVM; the absence of gradient boosting baselines — particularly XGBoost and LightGBM, which have demonstrated competitive or superior performance in several enterprise IS classification tasks [22], [25] — limits the completeness of the comparative landscape. Fifth, only the standard SMOTE variant was evaluated; alternative strategies such as Borderline-SMOTE, ADASYN, and cost-sensitive learning were not included, leaving open the question of whether a different imbalance correction method would alter the relative classifier rankings observed here [10], [33].

Building on the findings and limitations of this study, five directions for future research are recommended. First, the benchmark should be extended to include gradient boosting classifiers — specifically XGBoost and LightGBM — which have demonstrated strong performance under

imbalanced conditions [22], [25] and would clarify whether the RF dominance observed here persists when tree-boosting alternatives are included in the comparative set. Second, future studies should replicate this benchmark using primary organizational data obtained directly from enterprise IS deployments, as real-world datasets introduce label noise, temporal drift, and organizational context effects that public benchmarks cannot capture [13], [14]. Third, the extension of this framework to streaming and online enterprise IS environments — where models must adapt to distributional shift in real time — represents a practically urgent research direction, particularly for financial fraud detection and supply chain risk monitoring where concept drift is endemic [31]. Fourth, the SHAP analysis conducted here is correlational in nature; future work should explore causal interpretability frameworks that can distinguish genuine predictive drivers from confounded proxy features, particularly in the HRM and CRM domains where managerial interventions based on model attributions carry organizational consequences [27], [28]. Fifth, consistent with the recommendation of Manzoor et al. [23], future benchmarks in enterprise IS classification should incorporate profit-based and cost-sensitive evaluation metrics alongside standard classification measures, as the operational utility of a deployed model ultimately depends on the asymmetric costs of false positives and false negatives within each specific enterprise context.

V. CONCLUSION

This study presented a systematic comparative evaluation of Decision Tree, Random Forest, and Support Vector Machine classifiers across four enterprise IS classification tasks — employee attrition (HRM), customer churn (CRM), supplier delivery risk (SCM), and financial fraud detection — under a unified experimental framework incorporating SMOTE-based imbalance correction, grid search cross-validation, and SHAP explainability analysis. The central empirical finding is that Random Forest consistently achieves the highest F1-Score and AUC-ROC across all four domains, with SVM demonstrating competitive and occasionally superior precision in high-cost false-positive contexts such as HRM, and Decision Tree maintaining practical value through its interpretability advantage — particularly in near-balanced conditions where its F1 performance is statistically indistinguishable from SVM. A key cross-domain regularity established by this study is that imbalance ratio functions as a systematic moderator of classifier performance: SMOTE benefits scale inversely with IR, ranging from a marginal +0.78 pp F1 gain in the near-balanced SCM domain to a substantial +26.63 pp gain in the extreme-imbalance finance domain (IR = 0.17%), with the RF–DT and RF–SVM performance gaps correspondingly widening as imbalance severity increases. SHAP analysis of the bestperforming RF model per domain yielded actionable feature attributions directly applicable to enterprise decision-making: overtime exposure and job satisfaction in HRM, contract type and tenure in CRM, shipping mode

and order priority in SCM, and PCA component V14 alongside transaction amount in financial fraud detection.

The contributions of this study are fourfold. First, it provides the first rigorous multi-domain benchmark of DT, RF, and SVM under methodologically identical conditions spanning HRM, CRM, SCM, and financial enterprise IS contexts. Second, it empirically quantifies the interaction between class imbalance severity and SMOTE efficacy across a continuous IR range, offering a calibrated, data-driven basis for preprocessing decisions that single-domain benchmarks cannot provide. Third, the integration of SHAP explainability as a structural pipeline component — rather than an optional post-hoc step — produces feature attribution evidence validated simultaneously across four enterprise IS domains, strengthening interpretability claims beyond what any single-domain XAI study can establish. Fourth, the cross-domain empirical regularities identified here advance the enterprise IS classification literature toward a transferable, domain-agnostic evaluation standard applicable to practitioners and researchers selecting classifiers for new enterprise contexts. These findings collectively support the recommendation that enterprise IS stakeholders prioritize RF as a default classifier under moderate-to-severe class imbalance, consider SVM when precision constraints outweigh recall requirements, and retain DT when model transparency and computational efficiency are primary operational constraints. Future work should extend this framework to gradient boosting architectures, primary organizational datasets, streaming environments, and cost-sensitive evaluation metrics to further strengthen its practical and theoretical generalizability.

REFERENCES

[1] Y. Niu, L. Ying, J. Yang, M. Bao, and C. B. Sivaparthipan, "Organizational business intelligence and decision making using big data analytics," *Inf. Process. Manag.*, vol. 58, no. 6, p. 102725, Nov. 2021, doi: 10.1016/j.ipm.2021.102725.

[2] S. Ren, "Optimization of Enterprise Financial Management and Decision-Making Systems Based on Big Data," *Journal of Mathematics*, vol. 2022, no. 1, p. 1708506, Jan. 2022, doi: 10.1155/2022/1708506.

[3] A. Dogan and D. Birant, "Machine learning and data mining in manufacturing," *Expert Syst. Appl.*, vol. 166, p. 114060, Mar. 2021, doi: 10.1016/j.eswa.2020.114060.

[4] I. D. Mienye and N. Jere, "A Survey of Decision Trees: Concepts, Algorithms, and Applications," *IEEE Access*, vol. 12, pp. 86716–86727, 2024, doi: 10.1109/ACCESS.2024.3416838.

[5] Z. Sun, G. Wang, P. Li, H. Wang, M. Zhang, and X. Liang, "An improved random forest based on the classification accuracy and correlation measurement of decision trees," *Expert Syst. Appl.*, vol. 237, p. 121549, Mar. 2024, doi: 10.1016/j.eswa.2023.121549.

[6] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," *Neurocomputing*, vol. 408, pp. 189–215, Sep. 2020, doi: 10.1016/j.neucom.2019.10.118.

[7] Q. Li and J. Zhou, "A Comparative Analysis of Extreme Gradient Boosting, Decision Tree, Support Vector Machines, and Random Forest Algorithm in Data Analysis of

College Students' Psychological Health," *Informatica*, vol. 49, no. 15, pp. 127–134, Mar. 2025, doi: 10.31449/inf.v49i15.7004.

[8] L. Prado Osco *et al.*, "Forest Land Resource Information Acquisition with Sentinel-2 Image Utilizing Support Vector Machine, K-Nearest Neighbor, Random Forest, Decision Trees and Multi-Layer Perceptron," *Forests 2023, Vol. 14, Page 254*, vol. 14, no. 2, p. 254, Jan. 2023, doi: 10.3390/f14020254.

[9] M. Khushi *et al.*, "A Comparative Performance Analysis of Data Resampling Methods on Imbalance Medical Data," *IEEE Access*, vol. 9, pp. 109960–109975, 2021, doi: 10.1109/ACCESS.2021.3102399.

[10] M. Mujahid *et al.*, "Data oversampling and imbalanced datasets: an investigation of performance for machine learning and feature engineering," *Journal of Big Data 2024 11:1*, vol. 11, no. 1, pp. 87-, Jun. 2024, doi: 10.1186/s40537-024-00943-4.

[11] R. Dwivedi *et al.*, "Explainable AI (XAI): Core Ideas, Techniques, and Solutions," *ACM Comput. Surv.*, vol. 55, no. 9, Sep. 2023, doi: 10.1145/3561048.

[12] J. Gerlach, P. Hoppe, S. Jagels, L. Licker, and M. H. Breitner, "Decision support for efficient XAI services - A morphological analysis, business model archetypes, and a decision tree," *Electronic Markets 2022 32:4*, vol. 32, no. 4, pp. 2139–2158, Nov. 2022, doi: 10.1007/s12525-022-00603-6.

[13] V. Plotnikova, M. Dumas, and F. Milani, "Adaptations of data mining methodologies: a systematic literature review," *PeerJ Comput. Sci.*, vol. 6, pp. 1–43, 2020, doi: 10.7717/PEERJ-CS.267.

[14] F. Martinez-Plumed *et al.*, "CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 8, pp. 3048–3061, Dec. 2019, doi: 10.1109/TKDE.2019.2962680.

[15] Z. N. Jawad and V. Balázs, "Machine learning-driven optimization of enterprise resource planning (ERP) systems: a comprehensive review," *Beni-Suef University Journal of Basic and Applied Sciences 2023 13:1*, vol. 13, no. 1, pp. 4-, Jan. 2024, doi: 10.1186/s43088-023-00460-y.

[16] R. Panigrahi, N. Bele, P. K. Panigrahi, and B. B. Gupta, "Features level sentiment mining in enterprise systems from informal text corpus using machine learning techniques," *Enterp. Inf. Syst.*, vol. 18, no. 5, May 2024, doi: 10.1080/17517575.2024.2328186.

[17] D. Chung, J. Yun, J. Lee, and Y. Jeon, "Predictive model of employee attrition based on stacking ensemble learning," *Expert Syst. Appl.*, vol. 215, p. 119364, Apr. 2023, doi: 10.1016/j.eswa.2022.119364.

[18] B. Gaye, D. Zhang, and A. Wulamu, "Improvement of Support Vector Machine Algorithm in Big Data Background," *Math. Probl. Eng.*, vol. 2021, no. 1, p. 5594899, Jan. 2021, doi: 10.1155/2021/5594899.

[19] C. W. Lee, M. W. Fu, C. C. Wang, and M. I. Azis, "Evaluating Machine Learning Algorithms for Financial Fraud Detection: Insights from Indonesia," *Mathematics 2025, Vol. 13, Page 600*, vol. 13, no. 4, p. 600, Feb. 2025, doi: 10.3390/math13040600.

[20] N. Rezki and M. Mansouri, "Machine Learning for Proactive Supply Chain Risk Management: Predicting Delays and Enhancing Operational Efficiency," *Management Systems in Production Engineering*, vol. 32, no. 3, pp. 345–356, Sep. 2024, doi: 10.2478/mspe-2024-0033.

[21] P. Lalwani, M. K. Mishra, J. S. Chadha, and P. Sethi, "Customer churn prediction system: a machine learning approach," *Computing 2021 104:2*, vol. 104, no. 2, pp. 271–294, Feb. 2021, doi: 10.1007/s00607-021-00908-y.

[22] M. Imani, A. Beikmohammadi, and H. R. Arabnia, "Comprehensive Analysis of Random Forest and XGBoost Performance with SMOTE, ADASYN, and GNUS Under Varying Imbalance Levels," *Technologies 2025, Vol. 13, Page 88*, vol. 13, no. 3, p. 88, Feb. 2025, doi: 10.3390/technologies13030088.

[23] A. Manzoor, M. Atif Qureshi, E. Kidney, and L. Longo, "A Review on Machine Learning Methods for Customer Churn Prediction and Recommendations for Business Practitioners," *IEEE Access*, vol. 12, pp. 70434–70463, 2024, doi: 10.1109/ACCESS.2024.3402092.

[24] Y. Rimal, N. Sharma, and A. Alsadoon, "The accuracy of machine learning models relies on hyperparameter tuning: student result classification using random forest, randomized search, grid search, bayesian, genetic, and optuna algorithms," *Multimedia Tools and Applications 2024 83:30*, vol. 83, no. 30, pp. 74349–74364, Feb. 2024, doi: 10.1007/s11042-024-18426-2.

[25] J. Jin and Y. Zhang, "The analysis of fraud detection in financial market under machine learning," *Scientific Reports 2025 15:1*, vol. 15, no. 1, pp. 29959-, Aug. 2025, doi: 10.1038/s41598-025-15783-2.

[26] M. Liao, W. Jiao, and J. Zhang, "Research on Trade Credit Risk Assessment for Foreign Trade Enterprises Based on Explainable Machine Learning," *Information 2025, Vol. 16, Page 831*, vol. 16, no. 10, p. 831, Sep. 2025, doi: 10.3390/info16100831.

[27] K. Konar, S. Das, S. Das, and S. Misra, "Employee Attrition Prediction Using Bayesian Optimized Stacked Ensemble Learning and Explainable AI," *SN Computer Science 2025 6:6*, vol. 6, no. 6, pp. 672-, Jul. 2025, doi: 10.1007/s42979-025-04204-w.

[28] A. Raza, K. Munir, M. Almutairi, F. Younas, and M. M. S. Fareed, "Predicting Employee Attrition Using Machine Learning Approaches," *Applied Sciences 2022, Vol. 12, Page 6424*, vol. 12, no. 13, p. 6424, Jun. 2022, doi: 10.3390/app12136424.

[29] K. G. Al-Hashedi and P. Magalingam, "Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019," *Comput. Sci. Rev.*, vol. 40, p. 100402, May 2021, doi: 10.1016/j.cosrev.2021.100402.

[30] X. Liu, G. Xia, X. Zhang, W. Ma, and C. Yu, "Customer churn prediction model based on hybrid neural networks," *Scientific Reports 2024 14:1*, vol. 14, no. 1, pp. 30707-, Dec. 2024, doi: 10.1038/s41598-024-79603-9.

[31] Y. Lei, H. Qiaoming, and Z. Tong, "Research on Supply Chain Financial Risk Prevention Based on Machine Learning," *Comput. Intell. Neurosci.*, vol. 2023, no. 1, p. 6531154, Jan. 2023, doi: 10.1155/2023/6531154.

[32] F. Pedregosa FABIANPEDREGOSA *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011, Accessed: Mar. 12, 2026. [Online]. Available: http://jmlr.org/papers/v12/pedregosa11a.html

[33] D. Elreedy *et al.*, "A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning," *Machine Learning 2023 113:7*, vol. 113, no. 7, pp. 4903–4923, Jan. 2023, doi: 10.1007/s10994-022-06296-4.

[34] H. Wang, Q. Liang, J. T. Hancock, and T. M. Khoshgoftaar, "Feature selection strategies: a comparative analysis of SHAP-value and importance-based methods," *Journal of Big Data 2024 11:1*, vol. 11, no. 1, pp. 44-, Mar. 2024, doi: 10.1186/s40537-024-00905-w.

[35] S. Najafi-Zangeneh, N. Shams-Gharneh, A. Arjomandi-Nezhad, and S. H. Zolfani, "An Improved Machine Learning-Based Employees Attrition Prediction Framework with Emphasis on Feature Selection," *Mathematics 2021, Vol. 9, Page 1226*, vol. 9, no. 11, p. 1226, May 2021, doi: 10.3390/math9111226.

[36] S. Wu, W. C. Yau, T. S. Ong, and S. C. Chong, "Integrated Churn Prediction and Customer Segmentation Framework for Telco Business," *IEEE Access*, vol. 9, pp. 62118–62136, 2021, doi: 10.1109/ACCESS.2021.3073776.

[37] K. R. Ahmed, M. E. Ansari, M. N. Ahsan, A. Rohan, M. B. Uddin, and M. A. H. Rivin, "Deep learning framework for interpretable supply chain forecasting using SOM ANN and SHAP," *Scientific Reports 2025 15:1*, vol. 15, no. 1, pp. 26355-, Jul. 2025, doi: 10.1038/s41598-025-11510-z.

[38] E. Ileberi, Y. Sun, and Z. Wang, "Performance Evaluation of Machine Learning Methods for Credit Card Fraud Detection Using SMOTE and AdaBoost," *IEEE Access*, vol. 9, pp. 165286–165294, 2021, doi: 10.1109/ACCESS.2021.3134330.

[39] F. K. Alarfaj, I. Malik, H. U. Khan, N. Almusallam, M. Ramzan, and M. Ahmed, "Credit Card Fraud Detection Using State-of-the-Art Machine Learning and Deep Learning Algorithms," *IEEE Access*, vol. 10, pp. 39700–39715, 2022, doi: 10.1109/ACCESS.2022.3166891.

[40] J. G. Brandão *et al.*, "Optimization of machine learning models for sentiment analysis in social media," *Inf. Sci. (N. Y).*, vol. 694, p. 121704, Mar. 2025, doi: 10.1016/j.ins.2024.121704.

[41] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics 2019 21:1*, vol. 21, no. 1, pp. 6-, Jan. 2020, doi: 10.1186/s12864-019-6413-7.

[42] M. M. Ghiasi, S. Zendehboudi, and A. A. Mohsenipour, "Decision tree-based diagnosis of coronary artery disease: CART model," *Comput. Methods Programs Biomed.*, vol. 192, p. 105400, Aug. 2020, doi: 10.1016/j.cmpb.2020.105400.

[43] K. L. Du, B. Jiang, J. Lu, J. Hua, and M. N. S. Swamy, "Exploring Kernel Machines and Support Vector Machines: Principles, Techniques, and Future Directions," *Mathematics 2024, Vol. 12, Page 3935*, vol. 12, no. 24, p. 3935, Dec. 2024, doi: 10.3390/math12243935.

[44] Y. Nohara, K. Matsumoto, H. Soejima, and N. Nakashima, "Explanation of machine learning models using shapley additive explanation and application for real data in hospital," *Comput. Methods Programs Biomed.*, vol. 214, p. 106584, Feb. 2022, doi: 10.1016/j.cmpb.2021.106584.