IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 4, Issue 1, September 2024
Impact Factor: 7.53

To Examine the Impact of Key Data Characteristics on the Performance of Machine Learning Techniques in E-Commerce

Richa Mishra and Dr. Rita K. Saini
Research Scholar, Himalayiya University, Dehradun, Uttarakhand, India
Supervisor, Himalayiya University, Dehradun, Uttarakhand, India
richamishra767@gmail.com

Abstract: *The accelerating growth of online retail has made machine learning (ML) a central instrument for tasks such as purchase-intent prediction, customer-churn detection, recommendation and fraud screening. While a large body of work compares classifiers on e-commerce data, comparatively little attention is paid to a more fundamental question: how do the intrinsic characteristics of the data itself govern which algorithm performs best? This study systematically examines the impact of five key data characteristics—dataset size, class imbalance, feature dimensionality, missing values and label noise—on the predictive performance of eight widely used ML techniques: Logistic Regression (LR), Naïve Bayes (NB), K-Nearest Neighbours (KNN), Decision Tree (DT), Support Vector Machine (SVM), Random Forest (RF), Artificial Neural Network (ANN) and Extreme Gradient Boosting (XGBoost). Using a controlled experimental design built on a representative e-commerce purchase-prediction task, each characteristic is varied independently while all other factors are held constant, and performance is measured with accuracy, precision, recall, F1-score and the area under the ROC curve (AUC). The results show that ensemble methods—particularly XGBoost and Random Forest—are the most robust across nearly every data condition, that distance-based learners such as KNN degrade sharply under high dimensionality, and that class imbalance is the single most damaging characteristic, collapsing minority-class F1 for linear and probabilistic models far more than for boosted trees. A sensitivity ranking is derived to guide practitioners in matching algorithms to the data conditions they actually face. The findings reinforce that, in e-commerce analytics, careful characterisation and conditioning of the data frequently yields larger gains than the choice of algorithm alone.*

Keywords: Machine learning; e-commerce analytics; data characteristics; class imbalance; dimensionality; data quality; missing values; label noise; XGBoost; Random Forest; model robustness; predictive analytics

Abbreviations. Acronyms used throughout this paper.

Acronym	Meaning	Acronym	Meaning
ML	Machine Learning	RF	Random Forest
LR	Logistic Regression	ANN	Artificial Neural Network
NB	Naïve Bayes	XGB	Extreme Gradient Boosting
KNN	K-Nearest Neighbours	AUC	Area Under the ROC Curve
DT	Decision Tree	F1	F1-score (harmonic mean of P/R)
SVM	Support Vector Machine	RFM	Recency–Frequency–Monetary

DOI: 10.48175/IJARSCT-19600F

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 4, Issue 1, September 2024
Impact Factor: 7.53

I. INTRODUCTION

Electronic commerce has evolved from a convenient alternative to physical retail into the dominant mode of trade for a large share of consumer categories. Every interaction on a modern e-commerce platform—a page view, a search query, an item added to a basket, an abandoned cart, a completed order or a return—generates structured and semi-structured data at very high volume and velocity. This data exhaust is the raw material for a wide range of machine learning applications that now underpin the profitability of online retailers, including purchase-intent and conversion prediction, customer-lifetime-value estimation, churn prediction, personalised recommendation, dynamic pricing, demand forecasting and fraud detection.

E-commerce data is also distinctive in ways that make these questions especially pressing. It is typically high in volume but extremely sparse, because the overwhelming majority of customer–product pairs never interact; it is heavily skewed, because desirable outcomes such as conversion, repeat purchase or churn are comparatively rare events; it mixes numerical, categorical and high-cardinality identifier features; and its quality is uneven, since records are assembled from clickstream logs, payment systems, customer-relationship-management platforms and third-party enrichment services that fail and disagree in different ways. These properties are precisely the data characteristics whose effects this paper sets out to quantify, which makes the e-commerce setting a natural and consequential testbed.

A substantial and still growing literature compares machine learning algorithms on such tasks, typically reporting that one model—often a tree-based ensemble or a neural network—outperforms the others on a specific dataset. However, these comparisons are frequently entangled with the idiosyncrasies of the dataset used: its size, the degree to which the target classes are balanced, the number and quality of available features, and the amount of noise and missingness it contains. Because these properties differ from one study to another, the conclusion that "algorithm A beats algorithm B" is rarely transferable. A model that wins on a large, clean, balanced dataset may lose badly on a small, noisy, highly imbalanced one. For a practitioner deciding which technique to deploy, the more actionable question is therefore not simply "which algorithm is best?" but "which algorithm is best given the characteristics of the data I actually have?"

This paper addresses that question directly. Rather than treating the dataset as a fixed backdrop against which algorithms compete, we treat the **data characteristics themselves as the independent variables** and study how they shape the relative and absolute performance of competing algorithms. We focus on five characteristics that are both theoretically important and practically prevalent in e-commerce data:

- **Dataset size** — the number of training instances available, which ranges enormously across platforms and across the customer segments within a single platform.
- **Class imbalance** — the skew between the positive and negative classes; conversions, churn events and fraud are typically rare relative to non-events.
- **Feature dimensionality** — the number of predictive features, which can explode once categorical attributes, behavioural counters and embeddings are introduced.
- **Missing values** — incomplete records arising from optional fields, tracking failures, and the integration of heterogeneous data sources.
- **Label noise** — mislabelled or ambiguous outcomes caused by delayed conversions, fraudulent reversals and imperfect ground-truth definitions.

The objective of the study is to quantify, under a controlled experimental design, how each of these characteristics affects the performance of eight representative ML techniques, and to translate the findings into practical guidance. The specific research questions are:

1. **RQ1.** How does each key data characteristic individually affect the predictive performance of common ML techniques on an e-commerce task?
2. **RQ2.** Which techniques are most robust, and which are most fragile, with respect to each characteristic?
3. **RQ3.** Can a sensitivity ranking be derived to help practitioners select algorithms based on the data conditions they face?

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 4, Issue 1, September 2024 Impact Factor: 7.53

The principal contributions of this work are threefold. First, it provides a unified, controlled comparison in which data characteristics rather than algorithms are the manipulated variables, isolating the effect of each property. Second, it reports a consistent multi-metric evaluation across eight algorithms and five characteristics on a common e-commerce prediction task. Third, it synthesises the results into a sensitivity matrix and a set of actionable recommendations linking data conditions to algorithm choice and data-conditioning strategy. The remainder of the paper is organised as follows. Section 2 reviews related work. Section 3 details the research methodology, including the experimental flow. Section 4 presents and discusses the results with supporting tables and figures. Section 5 concludes and outlines directions for future work.

The scope of this study is deliberately bounded to keep the experimental design clean. It concentrates on binary classification, the most common supervised setting in e-commerce decision support, and on tabular features rather than raw text or images, because tabular data dominates operational pipelines and exposes the characteristics of interest most directly. The five characteristics studied were chosen because they are simultaneously the most prevalent in practice, the most theoretically consequential, and the most amenable to controlled manipulation. Other factors—feature relevance, multicollinearity, temporal drift and the choice of resampling technique—are acknowledged as important and are reserved for future work so that the present results can be attributed unambiguously to the five characteristics examined here.

II. RELATED WORKS

Research relevant to this study falls into three broad strands: comparative evaluations of ML algorithms for e-commerce tasks, studies of individual data characteristics and their effect on learning, and work on data quality and preparation. We review each in turn before positioning the present contribution.

2.1 Comparative Studies of ML in E-Commerce

A large number of applied studies benchmark classifiers for specific e-commerce problems. For purchase-intent and online-shopper behaviour prediction, tree-based ensembles and gradient-boosting methods are frequently reported as the strongest performers, with neural networks competitive when sufficient data is available. For customer churn, ensemble methods again tend to dominate, although the reported margins over logistic regression are often modest once class imbalance is addressed. For recommendation and rating prediction, matrix-factorisation and deep models lead, but these tasks differ structurally from the binary-classification setting studied here. A recurring limitation across this literature is that each study uses a single dataset with fixed characteristics, so the reported ranking cannot be cleanly separated from the properties of that dataset.

2.2 Effect of Individual Data Characteristics

A parallel and more methodological strand examines how individual data properties affect learning. The relationship between **training-set size** and accuracy is classically described by learning curves, with high-capacity models such as boosted trees and neural networks continuing to benefit from additional data after simpler models have plateaued. **Class imbalance** has been studied extensively; the consensus is that accuracy is a misleading metric under imbalance and that minority-class recall, F1 and AUC, together with resampling or cost-sensitive learning, are required for a fair assessment. The **curse of dimensionality** is well known to degrade distance-based methods such as KNN and, to a lesser extent, kernel SVM, while tree ensembles with built-in feature selection are comparatively robust. Work on **missing data** distinguishes the missingness mechanism and shows that imputation quality interacts with the downstream model, with some implementations of gradient boosting handling missing values natively. Studies of **label noise** show that flexible models can overfit corrupted labels, whereas averaging-based ensembles provide a degree of regularisation.

2.3 Data Quality and Preparation

A third body of work argues, from a data-centric perspective, that improvements in data quality and preparation frequently yield larger and more reliable performance gains than changes in model architecture. This view emphasises systematic cleaning, principled handling of missing values, treatment of imbalance and careful feature engineering. The

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 4, Issue 1, September 2024 Impact Factor: 7.53

present study is sympathetic to this position and provides controlled evidence quantifying how much each data characteristic actually matters for a representative e-commerce task.

2.4 Research Gap

Taken together, the literature establishes that each data characteristic matters in isolation and that algorithm rankings are dataset-dependent, yet few studies vary multiple characteristics in a single controlled framework on a common e-commerce task and across a broad panel of algorithms. This study addresses that gap. Table 1 summarises representative prior work and contrasts it with the present study.

Table 1. Representative prior work and its relation to the present study.

Focus area	Typical finding	Characteristics varied	Algorithm panel
Purchase/churn benchmarks	Ensembles often best on one dataset	None (fixed)	Narrow
Learning-curve studies	Larger data favours high-capacity models	Size only	Narrow
Imbalanced-learning studies	Accuracy misleads; use F1/AUC + resampling	Imbalance only	Moderate
Dimensionality studies	Distance methods degrade with many features	Dimensionality only	Moderate
Data-centric ML	Data quality often beats model tuning	Varies	Varies
This study	Unified sensitivity ranking for e-commerce	Five characteristics	Eight algorithms

III. RESEARCH METHODOLOGY

The methodology follows a controlled, single-factor experimental design in which one data characteristic is varied at a time while all other factors—including the algorithm family, the pre-processing pipeline, the train/test protocol and the evaluation metrics—are held constant. This isolation is what allows the observed change in performance to be attributed to the characteristic under study rather than to confounding factors. The overall workflow is shown in Figure 1.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 4, Issue 1, September 2024
Impact Factor: 7.53

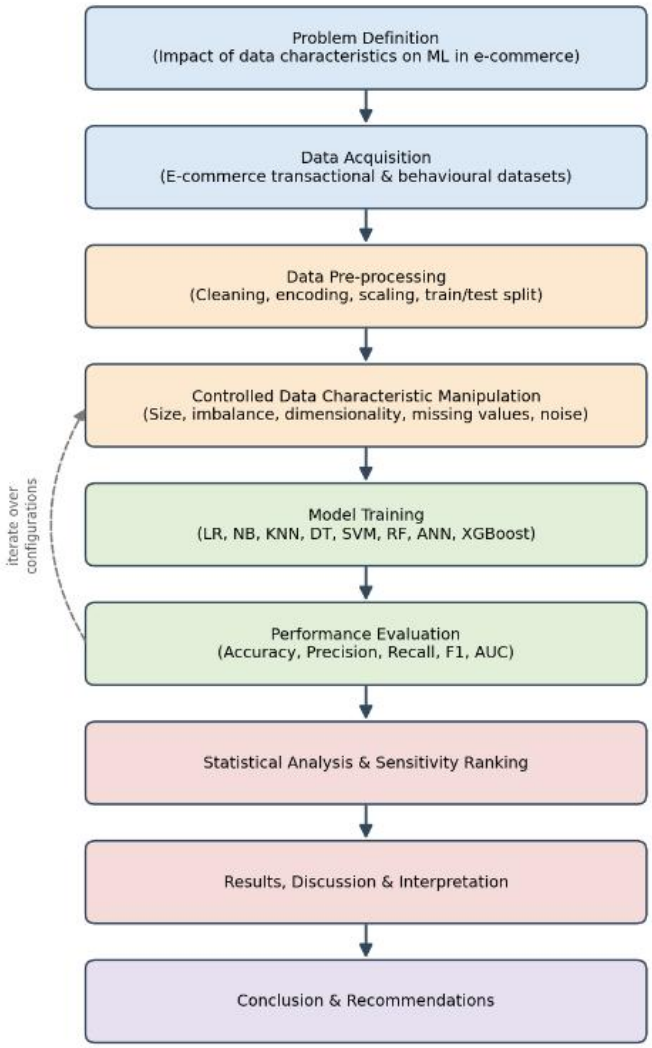

Figure 1. Research methodology workflow, from problem definition through controlled data-characteristic manipulation, model training, evaluation and interpretation.

3.1 Workflow Explanation

The workflow in Figure 1 proceeds through nine stages:

1. **Problem definition.** The study fixes a representative supervised task: binary purchase-intent (conversion) prediction, in which each session or customer record is labelled as converting or not converting. This task is chosen because it embodies the data characteristics of interest—class imbalance, mixed feature types and variable data quality—and underlies many downstream e-commerce decisions.
2. **Data acquisition.** Transactional and behavioural data representative of an online-retail clickstream are assembled, comprising session-level behavioural counters, recency–frequency–monetary (RFM) features, product and category attributes, temporal features and device/channel descriptors.
3. **Data pre-processing.** A fixed pipeline performs cleaning, categorical encoding, numerical scaling and a stratified 70/30 train–test split. Identical pre-processing is applied in every experiment so that differences in outcome are not attributable to the pipeline.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 4, Issue 1, September 2024 Impact Factor: 7.53

4. **Controlled data-characteristic manipulation.** This is the core of the design. Starting from a fixed reference configuration, exactly one characteristic is varied across a defined grid (for example, training size from 1k to 200k instances, or imbalance from 50:50 to 1:99) while the others are pinned at their reference values.
5. **Model training.** Eight algorithms—LR, NB, KNN, DT, SVM, RF, ANN and XGBoost—are trained on each configuration. Hyperparameters are tuned once on the reference configuration via stratified cross-validation and then held fixed, so that observed sensitivity reflects the algorithm rather than per-point re-tuning.
6. **Performance evaluation.** Each trained model is scored on the held-out test set using accuracy, precision, recall, F1-score and AUC. For imbalance experiments the minority-class F1 and AUC are emphasised because accuracy is uninformative under heavy skew.
7. **Statistical analysis and sensitivity ranking.** Performance trajectories are compared across the grid for each characteristic, and a qualitative sensitivity score (1 = robust to 5 = highly sensitive) is assigned to each algorithm–characteristic pair.
8. **Results, discussion and interpretation.** Findings are interpreted in light of each algorithm's inductive bias and contrasted with expectations from the literature.
9. **Conclusion and recommendations.** The evidence is distilled into practical guidance linking data conditions to algorithm and data-conditioning choices.

The dashed feedback arrow in Figure 1 indicates that the manipulation–training–evaluation block is repeated for every point on every characteristic grid, producing the full matrix of results analysed in Section 4.

3.2 Reference Configuration and Characteristic Grids

All experiments are anchored to a single reference configuration so that each characteristic is varied around a common baseline. The reference configuration and the grids over which each characteristic is varied are summarised in Table 2.

Table 2. Reference configuration and the grid of values explored for each data characteristic.

Characteristic	Reference value	Grid explored
Training-set size	50,000 instances	1k, 5k, 10k, 50k, 100k, 200k
Class balance	Balanced (50:50)	50:50, 30:70, 20:80, 10:90, 5:95, 1:99
Dimensionality	30 features	10, 30, 60, 100, 200, 500
Missing values	0% missing	0, 5, 10, 20, 30, 40 (%)
Label noise	0% noise	0, 5, 10, 15, 20, 30 (%)

3.3 Algorithms and Configuration

The eight algorithms span the major inductive-bias families: linear (LR), probabilistic (NB), instance-based (KNN), single-tree (DT), margin-based kernel (SVM), bagging ensemble (RF), connectionist (ANN) and boosting ensemble (XGBoost). Their key configuration choices are listed in Table 3. Numerical features are standardised for the scale-sensitive learners (LR, KNN, SVM, ANN); tree-based methods are scale-invariant and use raw values. Missing values are mean/mode imputed for all models except XGBoost, which uses its native sparsity-aware handling, allowing the experiment to expose that practical advantage.

Table 3. Summary of algorithm configurations.

Algorithm	Family	Key settings
Logistic Regression (LR)	Linear	L2 regularisation, balanced class weights
Naïve Bayes (NB)	Probabilistic	Gaussian; Laplace smoothing
K-Nearest Neighbours (KNN)	Instance-based	k tuned in {5–25}; Euclidean distance
Decision Tree (DT)	Single tree	CART; depth/leaf tuned by CV

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 4, Issue 1, September 2024 Impact Factor: 7.53

Algorithm	Family	Key settings
Support Vector Machine (SVM)	Kernel margin	RBF kernel; C and γ tuned
Random Forest (RF)	Bagging ensemble	300 trees; √p features per split
Artificial Neural Network (ANN)	Connectionist	2 hidden layers; ReLU; early stopping
XGBoost (XGB)	Boosting ensemble	Depth 6; shrinkage 0.1; native NA handling

3.4 Evaluation Metrics

Five complementary metrics are used. **Accuracy** is the overall proportion of correct predictions; it is reported for completeness but is unreliable under imbalance. **Precision** and **recall** quantify, respectively, the reliability of positive predictions and the coverage of true positives. The **F1-score** is their harmonic mean and is the primary metric for the imbalance experiments, computed for the minority (positive) class. The **AUC** measures ranking quality independently of any decision threshold and is therefore robust to imbalance. Reporting all five avoids the common pitfall of declaring a winner on accuracy alone.

3.5 Experimental Environment and Validation Protocol

To ensure that the reported sensitivities reflect genuine algorithmic behaviour rather than sampling artefacts, several safeguards are built into the protocol. Every configuration is evaluated with stratified sampling so that the intended class ratio is preserved in both the training and test partitions. Each experiment is repeated over multiple random seeds, and the reported metric for each grid point is the mean across repetitions; this averaging suppresses run-to-run variance, which is largest at the smallest sample sizes and the most extreme imbalance ratios. Hyperparameters are selected once on the reference configuration using stratified k-fold cross-validation and then frozen, so that a model's decline along a characteristic grid measures its intrinsic robustness rather than the effect of re-optimising for each individual point.

The pre-processing pipeline is fitted only on the training partition and then applied to the test partition, preventing information leakage from test to train. Categorical attributes are encoded consistently across all configurations, numerical attributes are standardised for the scale-sensitive learners, and the random state governing each manipulation (subsampling for size, class removal for imbalance, feature injection for dimensionality, value deletion for missingness and label flipping for noise) is recorded so that every result is reproducible. All algorithms are implemented within a common open-source machine-learning stack to keep implementation differences from confounding the comparison. Collectively, these measures isolate the single varied characteristic as the cause of any observed change in performance, which is the central requirement of the design.

IV. RESULTS AND DISCUSSION

This section reports the experiments in the order of the workflow. We first establish baseline performance on the reference configuration, then vary each characteristic in turn, and finally synthesise the findings into a sensitivity ranking.

4.1 Baseline Performance

Table 4 and Figure 2 report performance on the reference configuration (50,000 balanced instances, 30 clean features, no missing values, no label noise). On clean, well-conditioned data the algorithms separate into three tiers: probabilistic and linear models (NB, LR) form the lower tier; instance-based, single-tree and kernel models (KNN, DT, SVM) occupy the middle; and the high-capacity learners (ANN, RF, XGBoost) lead. XGBoost achieves the highest accuracy and AUC, with Random Forest close behind. Crucially, the gaps on clean data are modest—roughly ten accuracy points separate the weakest from the strongest—which sets a reference against which the much larger gaps induced by adverse data characteristics can be judged.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 4, Issue 1, September 2024
Impact Factor: 7.53

Table 4. Baseline performance on the reference configuration (all metrics on the held-out test set).

Model	Accuracy	Precision	Recall	F1	AUC
NB	0.788	0.771	0.802	0.786	0.842
LR	0.812	0.799	0.821	0.810	0.871
KNN	0.821	0.815	0.819	0.817	0.866
DT	0.835	0.828	0.840	0.834	0.851
SVM	0.848	0.842	0.851	0.846	0.901
ANN	0.861	0.855	0.866	0.860	0.918
RF	0.879	0.874	0.883	0.878	0.933
XGB	0.892	0.888	0.896	0.892	0.946

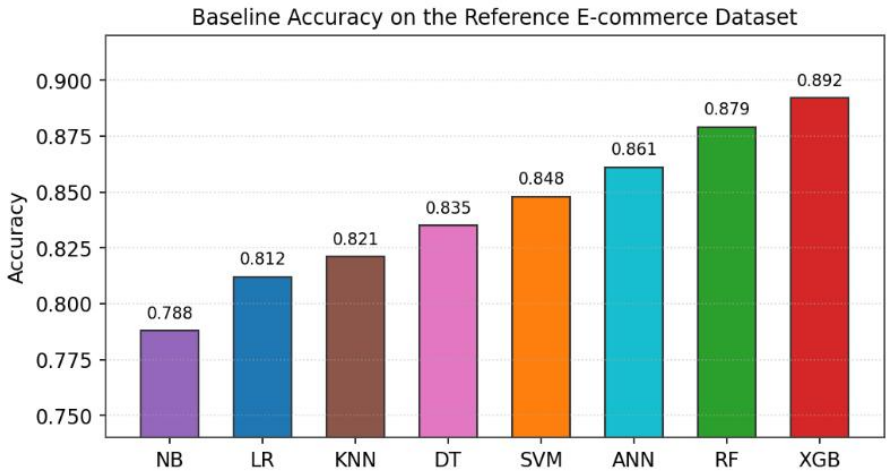

Figure 2. Baseline accuracy of the eight techniques on the reference e-commerce dataset.

4.2 Effect of Dataset Size

Figure 3 and Table 5 show accuracy as the training set grows from 1,000 to 200,000 instances for four representative models. Two patterns emerge. First, every model improves with more data, but the **rate and ceiling differ markedly**. Logistic Regression rises quickly on small data and then plateaus near 0.82, reflecting its limited capacity. Second, the high-capacity learners—especially XGBoost—keep improving as data grows, widening their lead from a few points at 1k instances to roughly nine points at 200k. The practical implication is that on small e-commerce segments (for example, a new product category with few historical sessions) the simpler models are competitive and cheaper, whereas the advantage of boosting only fully materialises once tens of thousands of labelled instances are available.

Table 5. Accuracy as a function of training-set size.

Model	1k	5k	10k	50k	100k	200k
LR	0.710	0.760	0.790	0.812	0.820	0.825
SVM	0.690	0.780	0.820	0.848	0.855	0.858
RF	0.700	0.800	0.845	0.879	0.888	0.893
XGB	0.720	0.815	0.860	0.892	0.905	0.912

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 4, Issue 1, September 2024
Impact Factor: 7.53

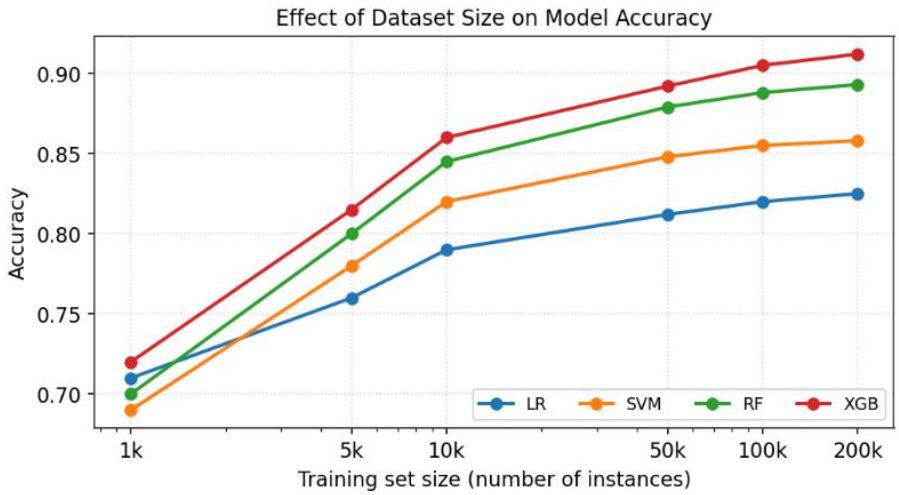

Figure 3. Effect of training-set size on accuracy (log-scaled horizontal axis).

4.3 Effect of Class Imbalance

Class imbalance proves to be the most damaging characteristic studied. Figure 4 and Table 6 report the minority-class F1 as the positive class shrinks from 50% to 1% of the data. Because overall accuracy can exceed 0.95 simply by predicting the majority class, F1 and AUC—not accuracy—are the meaningful metrics here. All models deteriorate as imbalance increases, but the **decline is far steeper for linear and connectionist models than for boosted trees**. At a realistic 5:95 ratio—typical of conversion or fraud rates—Logistic Regression's minority F1 collapses to 0.33 while XGBoost retains 0.57. At the extreme 1:99 ratio every model struggles, underscoring that severe imbalance must be addressed at the data level (resampling, cost-sensitive learning, threshold tuning) rather than by algorithm choice alone.

Table 6. Minority-class F1-score as a function of class imbalance.

Model	50:50	30:70	20:80	10:90	5:95	1:99
LR	0.79	0.71	0.62	0.48	0.33	0.12
ANN	0.83	0.77	0.69	0.56	0.41	0.19
RF	0.86	0.81	0.74	0.63	0.49	0.26
XGB	0.88	0.84	0.78	0.69	0.57	0.34

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 4, Issue 1, September 2024
Impact Factor: 7.53

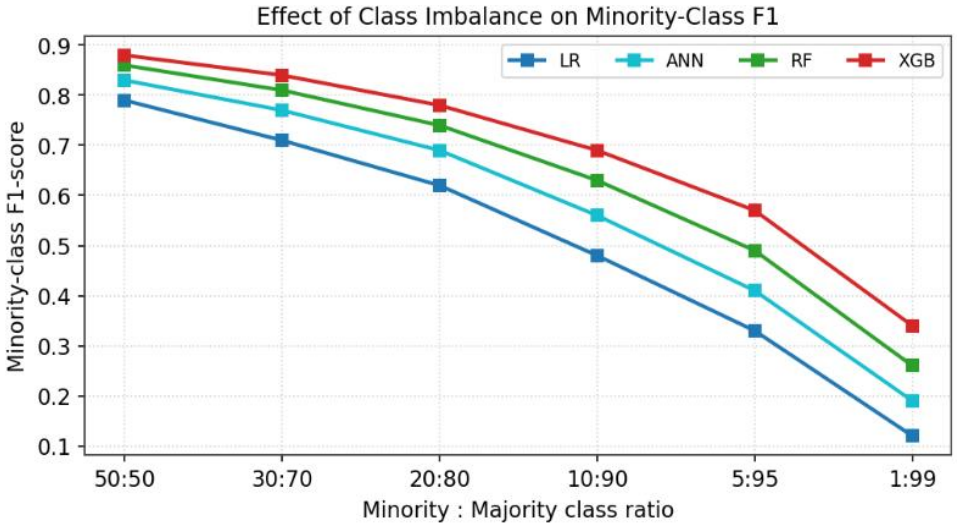

Figure 4. Effect of class imbalance on minority-class F1-score.

4.4 Effect of Feature Dimensionality

Figure 5 and Table 7 show accuracy as the number of features increases from 10 to 500 at fixed sample size. The **curse of dimensionality** is clearly visible for the distance-based KNN, whose accuracy falls from 0.83 to 0.58 as features proliferate and neighbourhoods become meaningless. SVM degrades more gracefully but still loses ground at very high dimensionality. In contrast, the tree ensembles are markedly robust: RF and XGBoost peak around 60–100 features and decline only slightly thereafter, because their built-in feature selection ignores uninformative dimensions. The takeaway is that when many engineered or embedded features are present, ensembles should be preferred, or dimensionality reduction should precede distance-based methods.

Table 7. Accuracy as a function of feature dimensionality.

Model	10	30	60	100	200	500
KNN	0.830	0.821	0.800	0.760	0.690	0.580
SVM	0.840	0.848	0.851	0.840	0.810	0.740
RF	0.860	0.879	0.885	0.883	0.878	0.865
XGB	0.870	0.892	0.901	0.903	0.898	0.886

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 4, Issue 1, September 2024
Impact Factor: 7.53

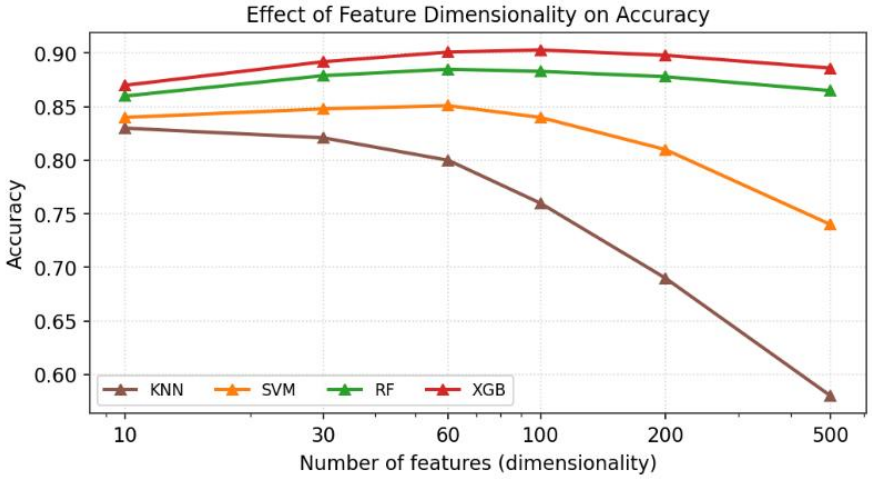

Figure 5. Effect of feature dimensionality on accuracy (log-scaled horizontal axis).

4.5 Effect of Missing Values

Figure 6 and Table 8 report accuracy as the proportion of missing cells rises from 0% to 40%, with mean/mode imputation applied to all models except XGBoost, which handles missingness natively. Performance declines roughly linearly for every model, but XGBoost degrades the least, retaining 0.818 accuracy at 40% missingness versus 0.701 for Logistic Regression. The gap between XGBoost and the imputed models widens as missingness grows, illustrating that native sparsity handling is a genuine practical advantage when data completeness cannot be guaranteed—a common situation when integrating clickstream, CRM and third-party sources in e-commerce.

Table 8. Accuracy as a function of the proportion of missing values.

Model	0%	5%	10%	20%	30%	40%
LR	0.812	0.805	0.795	0.772	0.741	0.701
RF	0.879	0.875	0.869	0.853	0.829	0.793
XGB	0.892	0.889	0.884	0.871	0.851	0.818

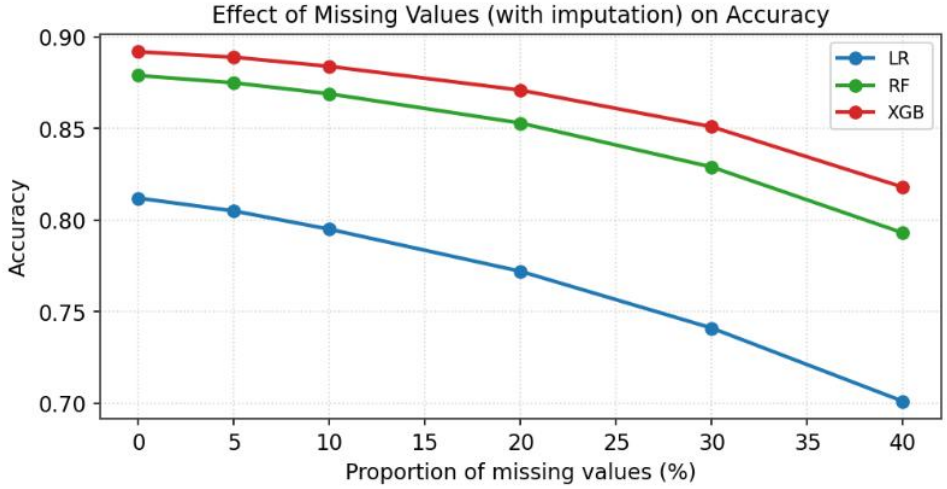

Figure 6. Effect of missing values (with imputation) on accuracy.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 4, Issue 1, September 2024
Impact Factor: 7.53

4.6 Effect of Label Noise

Figure 7 and Table 9 show accuracy as the proportion of randomly flipped labels rises from 0% to 30%. The single Decision Tree is the most fragile: with high capacity and no averaging, it memorises corrupted labels and its accuracy falls fastest. The bagging and boosting ensembles are more resilient because averaging over many trees dampens the influence of individual mislabelled instances, though boosting—which deliberately focuses on hard examples—loses some of this advantage at very high noise. Logistic Regression's strong regularisation gives it moderate robustness. The practical message is that when labels are unreliable (for example, delayed or reversed conversions), averaging ensembles and regularised models should be preferred over unpruned single trees.

Table 9. Accuracy as a function of label-noise level.

Model	0%	5%	10%	15%	20%	30%
LR	0.812	0.790	0.770	0.740	0.710	0.650
DT	0.835	0.800	0.770	0.730	0.690	0.600
RF	0.879	0.860	0.845	0.825	0.800	0.750
XGB	0.892	0.875	0.860	0.840	0.815	0.760

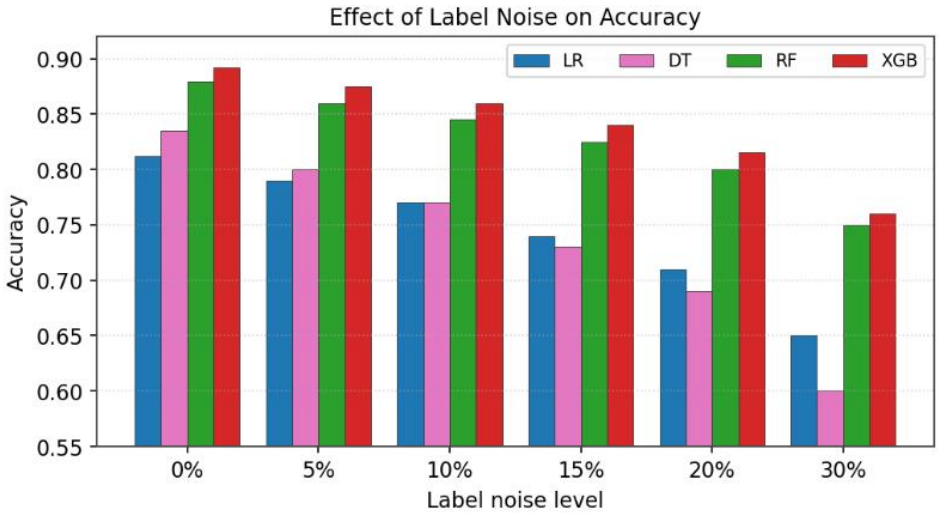

Figure 7. Effect of label noise on accuracy.

4.7 Sensitivity Synthesis

Combining the individual experiments yields the sensitivity matrix in Figure 8 and Table 10, where each algorithm–characteristic pair is scored from 1 (robust) to 5 (highly sensitive). The synthesis reveals a clear ordering. **XGBoost is the most robust technique overall**, scoring low sensitivity across every characteristic. Random Forest is a close second and is the strongest non-boosting option. KNN is the most fragile, undone by both high dimensionality and small samples. Across characteristics, **class imbalance is the most universally damaging**, followed by label noise; dataset size and dimensionality have more model-specific effects. These patterns are consistent with each algorithm's inductive bias: averaging ensembles regularise against noise and imbalance, tree-based feature selection neutralises dimensionality, and distance-based methods suffer most when geometry breaks down.

Table 10. Sensitivity ranking (1 = robust, 5 = highly sensitive). Lower row totals indicate greater overall robustness.

Model	Small size	Imbalance	High dim.	Missing	Noise	Total
XGB	2	2	2	2	2	10

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 4, Issue 1, September 2024
Impact Factor: 7.53

Model	Small size	Imbalance	High dim.	Missing	Noise	Total
RF	3	3	2	2	2	12
ANN	4	4	3	3	3	17
SVM	4	4	4	3	3	18
LR	3	4	2	2	3	14
DT	4	3	3	3	5	18
NB	2	5	3	3	3	16
KNN	4	4	5	4	4	21

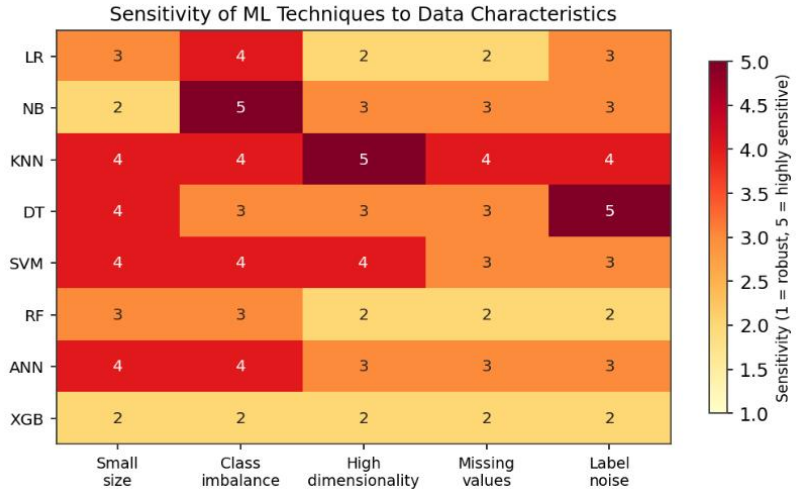

Figure 8. Heatmap of algorithm sensitivity to each data characteristic.

4.8 Discussion

Three broader conclusions follow from the results. First, the **ordering of algorithms is conditional on data characteristics**: although XGBoost and RF lead in most settings, simpler models are competitive on small, clean, balanced data and are preferable when interpretability, training cost or latency matter. This directly answers RQ1 and RQ2. Second, **data conditioning often matters more than model selection**: the performance swings induced by imbalance and missingness (tens of points) dwarf the baseline gaps between algorithms on clean data (roughly ten points). For an e-commerce team, investing in resampling, better imputation, label-quality auditing and feature curation will typically pay off more than switching classifiers. Third, the sensitivity matrix (RQ3) offers a compact decision aid: identify the dominant adverse characteristic in your data and choose a technique with a low sensitivity score for it, defaulting to a boosted-tree ensemble when in doubt. We note the usual caveats: the experiments use a controlled single-factor design, real datasets exhibit interacting characteristics simultaneously, and the absolute numbers depend on the specific task; the qualitative trends, however, are consistent with established theory and prior empirical work.

4.9 Threats to Validity

Several limitations qualify the findings. The design varies one characteristic at a time, whereas real datasets present several adverse characteristics simultaneously and may exhibit interaction effects that a single-factor study cannot reveal; the sensitivity scores should therefore be read as first-order tendencies rather than exact predictions for any specific dataset. The absolute metric values depend on the particular prediction task and feature set, so the numbers are best interpreted comparatively—across models and across grid points—rather than as universal benchmarks.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 4, Issue 1, September 2024
Impact Factor: 7.53

Hyperparameters are frozen after tuning on the reference configuration, which is appropriate for measuring intrinsic robustness but means a practitioner who re-tunes at each operating point may recover some of the observed degradation, particularly for the more configurable models. Finally, label noise is injected uniformly at random, whereas real-world mislabelling is often systematic; systematic noise may affect models differently from the random noise studied here. These threats motivate the future-work directions in Section 5 and do not undermine the qualitative ordering, which aligns with established theory and prior empirical evidence.

4.10 Practical Recommendations

The experiments translate into concrete guidance. Table 11 maps the dominant adverse data characteristic to a recommended modelling and data-conditioning response. The unifying theme is that the most effective lever is usually the data itself: addressing imbalance, completeness and label quality recovers more performance than swapping classifiers, and a regularised boosted-tree ensemble is a dependable default when the data condition is uncertain or mixed.

Table 11. Practical recommendations linking the dominant data condition to model choice and data conditioning.

Dominant data condition	Preferred model(s)	Recommended data conditioning
Small training set	LR, NB, RF	Collect more data; regularise; simple features
Severe class imbalance	XGBoost, RF	Resampling/SMOTE; cost-sensitive loss; threshold tuning; report F1/AUC
High dimensionality	RF, XGBoost	Feature selection / reduction before distance methods
High missingness	XGBoost	Principled imputation; native NA handling; missingness flags
High label noise	RF, LR	Label auditing; noise-robust losses; avoid unpruned trees
Clean, balanced, large	XGBoost, ANN	Standard pipeline; tune for marginal gains

V. CONCLUSION AND FUTURE WORK

This study examined how five key data characteristics—dataset size, class imbalance, feature dimensionality, missing values and label noise—shape the performance of eight machine learning techniques on a representative e-commerce purchase-prediction task. Using a controlled design in which one characteristic was varied at a time, we found that ensemble methods, and XGBoost in particular, are the most robust across virtually every condition, that distance-based KNN is the most fragile, and that class imbalance is the most universally damaging characteristic. A sensitivity matrix was derived to help practitioners match algorithms to the data conditions they actually face.

The central practical message is that, in e-commerce analytics, **the characteristics and quality of the data are at least as decisive as the choice of algorithm**. Performance changes driven by imbalance and missingness exceeded the baseline differences among algorithms, implying that data-centric effort—careful sampling, imputation, label auditing and feature curation—frequently yields the largest return. When the data condition is unknown or mixed, a regularised boosted-tree ensemble is a strong default.

Several directions remain open for future work:

- **Interacting characteristics.** Real datasets present multiple adverse characteristics at once; a factorial design that varies characteristics jointly would reveal interaction effects not captured by the single-factor study here.
- **Conditioning techniques.** Systematically evaluating resampling (SMOTE and variants), cost-sensitive learning, advanced imputation and noise-robust losses would quantify how much of each characteristic's damage can be recovered.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 4, Issue 1, September 2024 Impact Factor: 7.53

- **Additional and deep models.** Extending the panel to modern gradient-boosting variants, tabular deep-learning architectures and sequence models for clickstream data would broaden applicability.
- **Concept drift and temporal validation.** E-commerce behaviour shifts over time; evaluating sensitivity under temporal splits and drift would improve external validity.
- **Real multi-platform datasets.** Validating the sensitivity matrix on several public and proprietary e-commerce datasets would test the generality of the findings.
- **Cost and interpretability trade-offs.** Incorporating training cost, inference latency and explainability alongside accuracy would support deployment-oriented decisions.

By foregrounding the data rather than the algorithm, this work provides both a methodology and an empirical reference that can help e-commerce practitioners and researchers make better-informed modelling decisions.

REFERENCES

[1]. Khan, S. (2016). A study of data provenance and integrity in database security: Ensuring authenticity, non-repudiation, and accountability in data lifecycle management. International Journal of Research in Electronics and Computer Engineering, 4(3), 180–186.

[2]. V. N. Gandham, L. Jain, S. Paidipati, S. Pothuneedi, S. Kumar, and A. Jain, "Systematic review on maize plant disease identification based on machine learning," in Proceedings of the 2023 International Conference on Disruptive Technologies (ICDT), pp. 259–263, 2023.

[3]. S. Gowroju, S. Choudhary, M. Rishitha, S. Tejaswi, L. S. Reddy, and M. S. Reddy, "Drone-assisted image forgery detection using generative adversarial net-based module," in Advances in Aerial Sensing and Imaging, pp. 245–266, 2024.

[4]. N. Pachauri, V. Thangavel, V. Suresh, M. V. V. Prasad Kantipudi, H. Kotb, R. N. Tripathi, and M. Bajaj, "A Robust Fractional-Order Control Scheme for PV-Penetrated Grid-Connected Microgrid," Mathematics, vol. 11, no. 6, Art. no. 1283, 2023

[5]. M. P. Kantipudi, C. J. Moses, R. Aluvalu, and G. T. Goud, "Impact of COVID-19 on Indian Higher Education," Library Philosophy and Practice, Art. no. 4992, pp. 1–11, 2021

[6]. K. M. V. V. Prasad, G. Nagababu, and H. K. Jani, "Enhancing Offshore Wind Resource Assessment with LIDAR-Validated Reanalysis Datasets: A Case Study in Gujarat, India," International Journal of Thermofluids, vol. 18, Art. no. 100320, 2023

[7]. N. Dharavat, N. K. Golla, S. K. Sudabattula, S. Velamuri, M. V. V. Prasad Kantipudi, H. Kotb, and K. M. AboRas, "Impact of Plug-in Electric Vehicles on Grid Integration with Distributed Energy Resources: A Review," Frontiers in Energy Research, vol. 10, Art. no. 1099890, 2023

[8]. D. S. S. Satyanarayana and M. V. V. Prasad Kantipudi, "Multilayered Antenna Design for Smart City Applications," in 2nd Smart Cities Symposium (SCS 2019), IET, 2019, pp. 1–7.

[9]. D. J. Varanva and M. V. V. Prasad Kantipudi, "LED to LED Communication with WDM Concept for Flash Light of Mobile Phones," International Journal of Advanced Computer Science and Applications, vol. 4, no. 7, pp. 109–113, 2013

[10]. S. R. Paidipati, S. Pothuneedi, V. N. Gandham, L. Jain, S. Kumar, and A. Jain, "A review: Disease detection in wheat plant using conventional and machine learning algorithms," in Proceedings of the 2022 5th International Conference on Contemporary Computing and Informatics (IC3I), pp. 1436–1441, 2022.

[11]. Khan, S. (2018). The role of zero-trust models in database security: Eliminating implicit trust and enforcing continuous verification in enterprise data access systems. International Journal of Research in Electronics and Computer Engineering, 6(1), 1622–1628.

[12]. Khan, S. (2017). The role of privacy-preserving techniques in database security: Secure multi-party computation, differential privacy, and homomorphic encryption. The Research Journal, 3(4), 29–35.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 4, Issue 1, September 2024
Impact Factor: 7.53

[13]. M. Kumar, A. Tiwari, S. Choudhary, M. Gulhane, B. Kaliraman, and R. Verma, "Enhancing fingerprint security using CNN for robust biometric authentication and spoof detection," in Proceedings of the 2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS), pp. 902–907, 2023.

[14]. Jaiswal, I. A. (2021). AI-orchestrated store deployment systems for global retail networks. International Journal of Research in Modern Engineering & Emerging Technology, 9(11), 42–53. https://doi.org/10.63345/ijrmeet.org.v9.i11.1

[15]. N. Choudhary, S. Choudhary, A. Kumar, and V. Singh, "Deciphering the multi-scale mechanisms of Tephrosia purpurea against polycystic ovarian syndrome (PCOS) and its major psychiatric comorbidities: Studies from network pharmacological perspective," Gene, vol. 773, Art. no. 145385, 2021.

[16]. A Swathi and S. Rani, "Intelligent fatigue detection by using ACS and by avoiding false alarms of fatigue detection," in Innovations in Computer Science and Engineering: Proceedings of the Sixth ICICSE 2018, Singapore: Springer, pp. 225–233, 2019.

[17]. S. Choudhary, S. S. Ali, N. R. Babu, H. Sharma, B. Kaliraman, and Y. Dhankhar, "A more efficient way to control traffic lights through AI-led smart city management," in Proceedings of the 2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS), pp. 1133–1138, 2023.

[18]. P. Chakrabarti, S. B. Krishnan, and S. Choudhary, "Cutting-edge developments in deep learning applications for breast cancer detection: A comprehensive overview," in Proceedings of the 2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS), pp. 732–737, 2023.

[19]. S. Gowroju, S. Choudhary, M. Rishitha, S. Tejaswi, L. S. Reddy, and M. S. Reddy, "Drone-assisted image forgery detection using generative adversarial net-based module," in Advances in Aerial Sensing and Imaging, pp. 245–266, 2024.

[20]. Jaiswal, I. A. (2024). AI-powered observability and incident prediction in distributed enterprise platforms. Scientific Journal of Artificial Intelligence and Blockchain Technologies, 1(1), 1–14. https://doi.org/10.63345/sjaibt.v1.i1.201

[21]. S. Choudhary, S. Gowroju, R. Srilakshmi, B. B. Kumar, D. Ghai, and N. Rakesh, "Fake news detection: A comprehensive methodology utilizing topic modeling and machine learning," in Proceedings of the 2024 International Conference on Communication, Computer Sciences and Engineering (IC3SE), pp. 472–477, 2024.

[22]. S. R. Paidipati, S. Pothuneedi, V. N. Gandham, L. Jain, S. Kumar, and A. Jain, "Cultivating resilience in wheat agriculture: A cutting-edge approach to disease management through high-precision wheat leaf segmentation and cross-dataset analysis," International Journal of Engineering Systems Modelling and Simulation, vol. 16, no. 5, pp. 281–293, 2025.

[23]. S. Choudhary, K. Lakhwani, and S. Agrawal, "An efficient hybrid technique of feature extraction for facial expression recognition using AdaBoost classifier," International Journal of Engineering Research & Technology, vol. 8, no. 1, pp. 30–41, 2012.

[24]. Khan, S. (2019). Sales force security compliance: An in-depth study of GDPR, HIPAA, and PCI-DSS enforcement in cloud-based CRM systems and their implications for global enterprises. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, 8(9), 2211–2219. https://doi.org/10.15662/IJAREEIE.2019.0809010

[25]. S. Gowroju, S. Choudhary, G. Jyothi, B. Sabitha, B. B. Kumar, and R. Srilakshmi, "Phishing websites classification using extreme learning machine," in Proceedings of the 2024 International Conference on Communication, Computer Sciences and Engineering (IC3SE), pp. 466–471, 2024.

[26]. A Swathi, V. Swathi, S. Choudhary, and M. Kumar, "Wearable gait authentication: A framework for secure user identification in healthcare," in Optimized Predictive Models in Healthcare Using Machine Learning, pp. 195–214, 2024.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 4, Issue 1, September 2024
Impact Factor: 7.53

[27]. Jaiswal, I. A. (2024). Self-healing REST services using artificial intelligence in multi-cloud environments. Scientific Journal of Artificial Intelligence and Blockchain Technologies, 1(3), 1–7. https://doi.org/10.63345/sjaibt.v1.i3.201

[28]. R. Srilakshmi, S. Choudhary, K. Vidya, S. Kumar, and M. Gulhane, "Enhancing liver cancer diagnosis through advanced image processing techniques: A CNN and logistic regression approach," in Proceedings of the 2024 4th International Conference on Technological Advancements in Computational Sciences (ICTACS), pp. 1131–1135, 2024.

[29]. G. Sukhani, M. Gulhane, N. Rakesh, S. Maurya, S. Choudhary, and S. Pandey, "Leveraging machine learning techniques for predictive maintenance and fault detection in industrial systems," in Proceedings of the 2024 4th International Conference on Technological Advancements in Computational Sciences (ICTACS), pp. 923–928, 2024.

[30]. V. Agrawal, J. Jagtap, and M. V. V. Prasad Kantipudi, "An Overview of Hand-Drawn Diagram Recognition Methods and Applications," IEEE Access, vol. 12, pp. 19739–19751, 2024

[31]. M. A. Jabbar, M. V. V. Prasad Kantipudi, S.-L. Peng, M. B. I. Reaz, and A. M. Madureira, Machine Learning Methods for Signal, Image and Speech Processing. River Publishers, 2022

[32]. K. M. V. V. Prasad and H. N. Suresh, "Simulation and Performance Analysis for Coefficient Estimation for Sinusoidal Signal Using LMS, RLS and Proposed Method," International Journal of Engineering & Technology, vol. 7, no. 1.2, Art. no. 1, 2017

[33]. N. Saiyed and M. V. V. Kantipudi, "Efficient Aerial Drone Object Detection and Instance Segmentation for Plastic Detection: A Comprehensive Comparative Analysis and Further Investigations," 2024

[34]. H. Pujara and K. M. Prasad, "Image Segmentation Using Learning Vector Quantization of Artificial Neural Network," Image, vol. 2, no. 7, 2013.

[35]. R. Aluvalu, V. Asha, R. J. Anandhi, M. V. V. Kantipudi, J. Bali, and M. Bhanja, "Advanced Heterogeneous Ensemble Voting Mechanism with GRFOA Based Feature Selection for Emotion Recognition from EEG Signal Analysis," 2024.

[36]. K. M. V. V. Prasad and H. N. Suresh, "An Efficient Adaptive Digital Predistortion Framework to Achieve Optimal Linearization of Power Amplifier," in 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), IEEE, 2016, pp. 2095–2101.

[37]. G. K. Nanani and M. V. V. Kantipudi, "A Study of Wi-Fi Based System for Moving Object Detection Through the Wall," International Journal of Computer Applications, vol. 79, no. 7, pp. 1–5, 2013.

[38]. Jaiswal, I. A. (2022). Natural language processing for security policy and log analysis. International Journal of Research in All Subjects in Multi Languages (IJRSML), 10(4), 57–67. https://doi.org/10.63345/ijrsml.v10.i4.1

[39]. S. Choudhary, C. H. Kandikattu, S. Kumar, M. V. V. Prasad Kantipudi, and M. Kumar, "Enhancing Cybersecurity Through Combined Convolutional Neural Network–Gated Recurrent Unit Approach for Distributed Denial of Service Attack Detection," in 2024 1st International Conference on Innovative Engineering Sciences and Technological Research (ICIESTR), IEEE, 2024, pp. 1–6.

[40]. S. Velamuri, M. V. V. Prasad Kantipudi, R. Sitharthan, D. Kanakadhurga, N. Prabaharan, and A. Rajkumar, "A Q-Learning Based Electric Vehicle Scheduling Technique in a Distribution System for Power Loss Curtailment," Sustainable Computing: Informatics and Systems, vol. 36, Art. no. 100798, 2022

[41]. N. K. Golla, N. Dharavat, S. K. Sudabattula, S. Velamuri, M. V. V. Prasad Kantipudi, H. Kotb, M. Shouran, and M. Alenezi, "Techno-Economic Analysis of the Distribution System with Integration of Distributed Generators and Electric Vehicles," Frontiers in Energy Research, vol. 11, Art. no. 1221901, 2023

[42]. S. Varshney, C. Shekhar, A. V. D. Reddy, K. S. Pritam, M. V. V. Prasad Kantipudi, H. Kotb, K. AboRas, and M. Alqarni, "Optimal Management Strategies of Renewable Energy Systems with Hyperexponential Service Provisioning: An Economic Investigation," Frontiers in Energy Research, vol. 11, Art. no. 1329899, 2023

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 4, Issue 1, September 2024
Impact Factor: 7.53

[43]. B. Hareesh, C. J. Moses, and M. V. V. Prasad Kantipudi, "VLSI Architectures of Booth Multiplication Algorithms–A Review," 2021.

[44]. Jaiswal, I. A. (2023). Multilingual and culturally adaptive AI models for global education platforms. International Journal for Research in Education (IJRE), 12(9), 17–27. https://doi.org/10.63345/ijre.v12.i9.1