Received: 05 Mar 2025 | Revised: 18 Aug 2025 | Accepted: 01 Dec 2025 | Published: 30 Dec 2025
Volume 3(2) Pages 78-92 DOI: https://doi.org/ 10.56279/jicts.v3i2.288

JICTS
Journal of ICT Systems

Effect of Outliers on Performance of Machine Learning Models: An Empirical Analysis

Hussein Bakiri [a], Hadija Mbembati [b, 1]

[a]*Department of Computer Science, Institute of Finance Management, Dar es Salaam, Tanzania*
[b]*Department of Electronics and Telecommunications Engineering, University of Dar es Salaam, Dar es Salaam, Tanzania*

[1]Corresponding author
Email: hadija@udsm.ac.tz

Funding information

No funds were provided for this research.

Keywords

Machine learning models
Outlier effects
Outlier trends
Training accuracy

Abstract

Machine learning (ML) is a key component of artificial intelligence, capable of solving complex problems by mimicking natural behaviors. Existing studies on the outlier effect in machine learning typically examine fewer than 10 outliers per dataset, limiting analysis of how outlier trends influence model performance. This study addresses this gap through an empirical analysis using 200 outlier instances. The data were drawn from 105,192 electricity consumption records, of which 61,783 were identified as outliers. Four ML models were applied: Support Vector Regression (SVR), Artificial Neural Networks (ANN), Decision Trees (DT), and Random Forests (RF). Results indicate that the relationship between outliers and training accuracy is linear for SVR and RF. In contrast, ANN and DT exhibit nonlinear trends, especially when the outlier percentage falls below a critical threshold of 13.978%. In terms of performance, DT, RF, and SVR show lower sensitivity to outliers compared to ANN. These findings highlight the importance of careful model selection when handling datasets with significant outlier presence.

1. Introduction

Machine learning prediction models have become more popular in various applications to support key decision-making processes [1]. These models exploit the availability of big data to train the model and let the system learn the dataset pattern for prediction. The performance of these models depends on the ability of the model to learn and generalize from historical datasets, enabling an accurate prediction model. The quality and relevance of historical datasets are crucial in machine learning for making proper predictions [2]. However, historical datasets may have outliers, significantly affecting machine learning (ML) models' accuracy, training, and generalization [3].

Outliers are the data values that significantly deviate from other values in the dataset [4]. Several factors, including measurement, sampling, and human errors, can cause outliers [5]. Detecting outliers from datasets has broad applications in various fields, such as smart grids, medicine, and the manufacturing industry [6]. The most popular outlier detection method is visualization, which resembles Boxplot, Histogram, Scatter Plot, and interquartile range (IQR) methods [7]. Many outlier detection and cleansing models in different fields have been proposed in the literature. These models include those based on clustering, statistics, classification, and Artificial Intelligence (AI) [8].

Clustering-based outlier detection techniques have been proposed by Du et al. [9] to improve the performance of artificial neural network (ANN) prediction models. The datasets were clustered into three groups (unscreening, partial screening, and full screening), and the test results show that the prediction accuracy for full screening is superior over the other two groups of datasets. This observation indicates the outlier's effects on the ANN model's prediction accuracy. The effects of the outlier in the prediction models can be observed in different outlier percentages, as [9], [10].

The statistical-based outlier detection technique has been proposed to test the effect of outliers in different regression models [11]. The experimental results indicate that removing outliers significantly improve the accuracy of predictive models, as evidenced by reduced root mean square errors (MSE). Thus, outlier detection is critical in statistical analysis, and inappropriate handling of outliers may result in inaccurate conclusions [12].

Recently, classification outlier detection models, such as support vector machines and random forests, have been used to improve the performance of energy planning predictions. The model's comparative performance analysis by Ododo [13] indicates that the optimized support vector machine (SVM) has a high level of robustness, especially in handling datasets with outliers and missing data points. Subsequently, the optimized random forest classification for outlier detection improved the prediction accuracy [14].

One significant gap in the current literature is a limited focus on the effects of outliers on model training accuracy. While numerous studies address detecting and removing outliers to improve model accuracy, limited studies explore how outliers affect the model training accuracy and interpretability [13]. These studies have examined not more than 10 instances per dataset [10], [15], [16], [17], [18] . While this study has shed light on the effect of outliers in machine learning models, the outlier contamination percentage and scenarios are limited. Also, the studies do not consider the varying trend of the outlier effect on the training models. This narrow scope and lack of investigation of outlier variation trends may restrict the generalization of findings. They may obscure the actual impacts of outlier presence on machine learning model performance, particularly training accuracy.

This gap underscores the importance of research that explores the effects of outliers on model accuracy and interpretability, aiming to ensure that machine learning models remain trustworthy and informative in the presence of outliers. Given the prevalence of outliers in real-world data and the sensitivity of many ML models to such anomalies, this underexplored relationship presents a significant research gap worthy of further investigation. To bridge this gap, the authors aim to reveal the outlier effects on the training accuracy of ML models, such as ANN, Support Vector Regression (SVR), Random Forest, and Decision Tree Algorithms, based on 200 datasets. In particular, this work will have the theoretical and practical contributions as follows:

i. Establishing a comprehensive trend analysis of outlier effects

ii. Identifying a model-specific performance

Related Works

Outlier handling in machine learning models is quite challenging and can significantly affect the performance of the models. Researchers have presented several approaches to analyze the effects of outliers on the performance of ML models. For example, Ansari et al. [17] investigate the effects of outliers in the ML models. The study reveals that some modes, such as Multilayer Perceptron (MLP) models, exhibit the highest sensitivity to outliers, with a deterioration rate of 54.04%. The author tested only two instances, with and without an outlier. Another study uses the same technique of testing two instances and shows significant effects in model performance before outlier handling, with an accuracy decrease of 25% [18].

The recent study analyzed the effects of outliers using simulated data with a 5% contamination rate [16]. The results show that the model suffers from a performance decline and is highly sensitive to outliers. The authors recommend that future studies consider different scenarios of data contamination. The study by Mohammed et al. [15] tested three scenarios to analyze the effects of outlier data on machine learning models, including Support Vector Machine (SVM) and Neural Network (NN) models: polluted training set, polluted test set, and polluted training and test sets. The result indicates that the outlier can affect the performance of ML models by approximately 50%. Despite these efforts, further evaluation of the effects of outliers on ML models is required, considering different scenarios of dirty training datasets to deduce outlier performance trends and capture better generalizability.

2. Method

This study applies a mixed-methods approach to investigate how outliers affect the performance and accuracy of machine learning models. Four models were chosen: ANN, RF, DT, and SVR. The combination of these models provides a balanced comparison of algorithms, ranging from high-variance learners (DT) to ensemble-based models (RF) and nonlinear predictors (ANN and SVR). This diversity ensures a more generalizable understanding of how outlier contamination influences model performance [19]. Also, the study applies a comparative experimental research methodology by using a range of methods, including controlled data perturbation, 5-fold cross-validation, Ordinary Least Squares (OLS) regression, Interquartile Range (IQR), and visualization techniques such as line plots.

2.1 Controlled Data Perturbation

This method systematically injects synthetic outliers into the dataset at controlled percentages (e.g., 0% to 50%) to simulate real-world data contamination [19]. Controlled perturbation allows for a clear understanding of how varying noise levels affect model accuracy in a reproducible way. This technique has been widely adopted in robustness testing of machine learning algorithms [20]

2.2 Five-Fold Cross-Validation

To evaluate model performance reliably, the study used five-fold cross-validation, a standard resampling technique that reduces overfitting and improves generalization estimates [21]. This method splits the data into five equal parts, iteratively training on four folds and testing the remaining one, thereby averaging the results for better accuracy estimation [22]. It is advantageous when working with limited data and comparing multiple models under the same conditions [23].

2.3 Ordinary Least Squares (OLS) Regression

OLS was applied to model and evaluate the linear relationships between the percentage of outliers and the models' performance metrics. OLS is a foundational statistical method for estimating the parameters in linear regression models and is widely used in model performance analysis due to its interpretability and computational efficiency [24].

2.4 Interquartile Range (IQR) for Outlier Detection

The IQR method was utilized to identify and analyze natural outliers in the dataset before synthetic perturbation [26]. This method identifies outliers as values below the first quartile (Q1) minus 1.5 times the IQR or above the third quartile (Q3), as shown in equation 1. It is robust to non-normal data and is commonly used in exploratory data analysis [27], making it suitable for datasets with unknown distributions. Generally, these methods were chosen due to their effectiveness in analyzing model behavior under noise, their statistical rigor, and their wide acceptance in empirical machine learning and statistical research.

The point X considered an outlier if:

$$X < Q1 - 1.5 * IQR \; or \; X > Q3 + 1.5 * IQR \quad (1)$$

2.5 Conceptual design of the proposed outlier effect analysis

Figure 1 illustrates the systematic approach to investigating outlier analysis in model training and performance evaluation. The process began with data preprocessing to ensure data quality and consistency. The study performed an outlier analysis to identify anomalous data points after preprocessing. The processed data was then used to train the model using various outlier definitions based on different thresholds for defining outlier. After model training, the study assessed the effect of the outlier definitions on model performance.Based on this assessment, the outlier definition is adjusted, and the model retraining process is repeated until an optimal setting is reached.

2.6 Data Description and Preprocessing

This research used quantitative data characterized by continuous trends. The data were collected from the Tanzania Electricity Supply Company (TANESCO). Such data included electricity consumption records (load in kWh) collected using an Automated Meter Reading (AMR) device connected to the transformer.

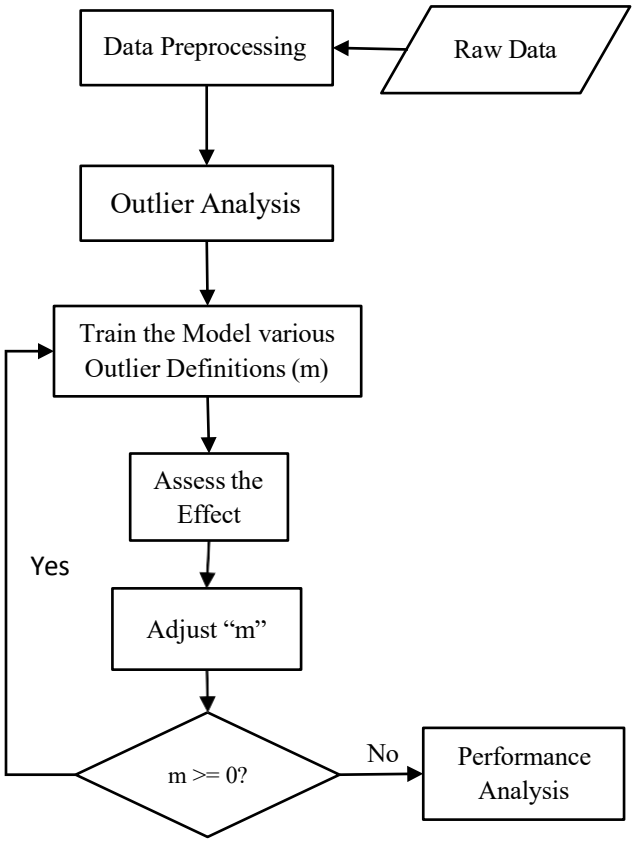


Figure 1. Proposed conceptual model of outlier sensitivity analysis.

This device was connected to one of the distribution transformers in the Kinondoni North area. It comprised twenty-minute interval readings in kWh recorded for four years. Another type of data used in this research was the time of day and day type. The total data points collected over the four years was 105,192

Each data entry includes a timestamp and the total energy consumed by all residential consumers connected to the transformer at that moment. The residential area under observation is characterized by medium electricity consumption, and the load pattern is primarily influenced by typical household activities, such as lighting, cooking, and the use of appliances.

Several preprocessing steps were applied to ensure the data's quality and consistency. These steps included outlier detection and time normalization. Outlier detection was accomplished using the Interquartile Range (IQR) method due to its robustness and ease of use. IQR makes it ideal for univariate outlier detection tasks [7]. After applying this method, the output was the number and percentage of outliers, determined based on the following formula in equation 2:

$$IQR = Q3 - Q1 \qquad (2)$$

Where:

$Q3$ = Third Quartile

$Q1$ = First Quartile

Furthermore, the lower and upper bounds are calculated as follows:

$$Lower\ Bound = Q1 - 1.5 * IQR$$

$$Upper\ Bound = Q3 - 1.5 * IQR$$

To normalize the time and day values in the dataset while preserving their cyclical nature, we applied cyclical encoding using sine and cosine transformations in equations 3 and 4. Features, such as hour and day of the week, were extracted from the timestamp and transformed accordingly. This method allows machine learning models to capture the periodic patterns in electricity consumption better, as it prevents discontinuities inherent in linear encodings [28].

$$x_{sin} = \sin(2\pi * x/p) \qquad (3)$$

$$\mathrm{xcos} = \cos(2\pi * \mathrm{x/p}) \qquad (4)$$

Where:

x: the original time value (e.g., hour, day)

p: the period of the cycle (e.g., 24 hours in a day, 7 days in a week)

2.7. Model Selection and Training

Four models were selected for performance comparison with the effect of outliers: ANN, Random Forest, Decision Trees, and SVR. These methods are popular and widely used due to their flexibility, broad applicability and versatility [29]. Furthermore, ANN, SVR, Decision Tree, and Random Forest models are chosen to represent a variety of algorithmic behaviors in response to outliers, ranging from simple rule-based and ensemble techniques to complex nonlinear models [30]. While Random Forest mitigates outlier effects through aggregation, SVR offers robustness via its margin-based approach, and ANN and Decision Trees are known to be sensitive to outliers [31]. Focusing on these well-established models simplifies the analysis of outlier effects by avoiding the introduction of less common algorithms.

In this study, the four models were trained with data containing the percentage of outliers detected (m). The training process iterated with a decrease in the outlier percentage. The value of "m" was decremented to obtain enough steps (i.e., 200) for analyzing the accuracy-outlier trends. The output of the outlier sensitivity analysis process is the percentage of outliers, accuracy, and performance (in terms of R-squared (R^2) and training time) for the four trained models. Furthermore, correlation and Mean Squared Error (MSE) values were also measured in each iteration. These steps are depicted in the conceptual model given in Figure 1.

2.8 Analyzing Model Performance

At this step, descriptive statistics and visual

.

analysis were used to analyze the effect of outliers on model performance. Two aspects of model evaluation metrics were used at this step: loss (MSE) and accuracy (R^2). In that sense R^2 values were used for accuracy and Mean Squared Error (MSE) for loss comparisonsConcerning visual analysis, curves for training performance, learning, and training time were plotted and analyzed. In addition, Taylor diagrams were plotted to grasp a deeper understanding of the correlational analysis process. Additionally, after observing the nature of the relationship between outliers and the model's training performance of ML algorithms, the study also investigated its significance level using the OLS method.

3. Experimental Setup

After carrying out the preprocessing and outlier analysis processes, the percentage of outliers (m_0), numbers of outliers (n_0) and non-outliers (c) were, respectively, as follows: 10.43%; 21,782; and 94,210, out of a total record (t_0) of 105,192. The portion of non-outliers (i.e., 40,000) and outliers (i.e., 21,782) was randomly mixed to get a combination of new records t' (61,783). This mixing was essential to increase the number of outlier definition in the data by 50%.

Another essential step to determine the percentage of outliers, m', iteratively from the new outlier definition (n') was carried out. In each iteration, the number of outliers (n') was reduced by a small number p, and a model was trained based on the percentage of outliers in the data. This process continued until n' becomes zero. The value of p was heuristically formulated to obtain many iterations enough to study the outlier-effect trend. In this study, the value of p was set to 100. Finally, the performance analysis process took place.

Table 1 presents the pseudocode that gives the whole process of outlier sensitivity analysis.

4. Results

This section presents the research findings aimed at answering the stated research questions. The research questions sought to analyze the effect of outliers for four machine learning methods: Decision Tree, Random Forest, SVR, and ANN. The analysis is based on three aspects: training accuracy, error, and the time of a model. Furthermore, they intend to probe in-depth into the trend in which outliers affect the training accuracy of machine learning models. In addition, the section presents an in-depth discussion concerning the obtained results.

Significance of outlier effect on a model's accuracy

The study findings indicate that outliers affect the accuracy of a model significantly. This effect is shown by the Taylor diagrams (Figures 2(a), 2(b), 2(c), and 2(d)). The effect seems to be the same in all four models (ANN, SVR, Random Forest, and Decision Tree). The diagrams present two important aspects: correlation coefficients and standard deviation. 5 are continuously below 0.05 (around 0.025), significantly less than the observed 0.20. The measured percentage outlier's variability throughout the instances is represented by the observed standard.

Table 1. The process of outlier sensitivity analysis.

Step 1:	Mix a portion of non-outliers c with detectedoutliers (n_0) to form a new record (t')
Step 2:	Determine new percentage outliers (m')from (t')
Step 3:	Train an ML model with (m') and (t')
Step 4:	Assess the outlier effect based on m'
Step 5:	Reduce the number of outliers by p ($n' = n' - p$)
Step 6:	If ($n' >= 0$) then repeat steps 2to step 5
Step 7:	Else, stop

The standard deviations from Figures 2 through deviation, which is used as a benchmark to evaluate how well the model reproduces the data's distribution. The large discrepancy between the values obtained from the models and the observed ones suggests that the models substantially underestimate the variability, resulting in smoother predictions that do not account for the short-term fluctuations in the observations. Furthermore, all four models produced correlation coefficient values that are nearly 0.2 which shows a weak relationship with the observed data and nearly further suggests poor agreement and prediction ability with respect to percentage outlier variability and performance of a model. Another aspect of model comparison with regard to performance is based on R-squared values. These values measure the proportion of the variance in the target variable explained by the model's features. Figure 3 presents how R-squared values are affected by the number of outliers in the data. The Figure shows how badly the model tries to fit the data that contains outlier.

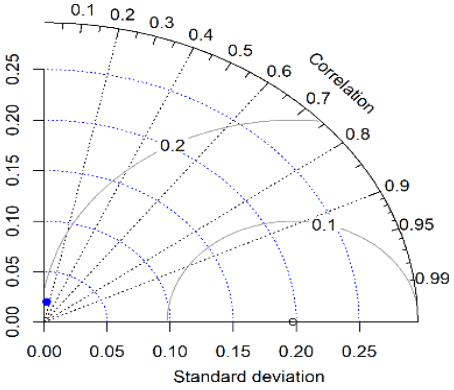


Figure 2(a). A Taylor diagram for an ANN model.

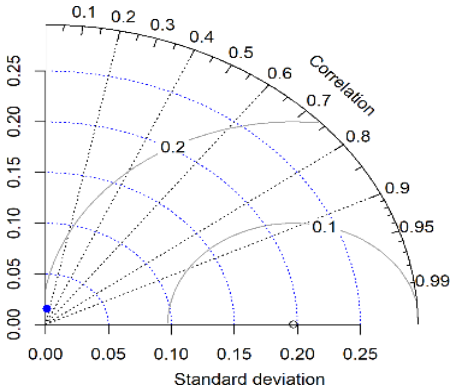


Figure 2(b). A Taylor diagram for a Decision Tree model

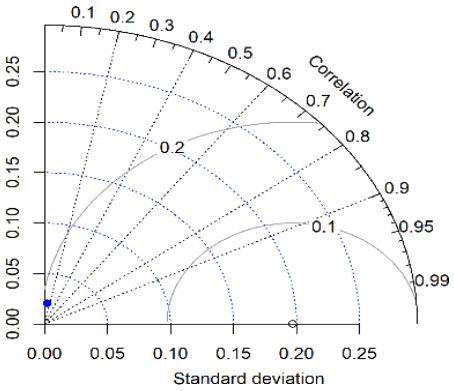


Figure 2(c). A taylor diagram for an SVR model

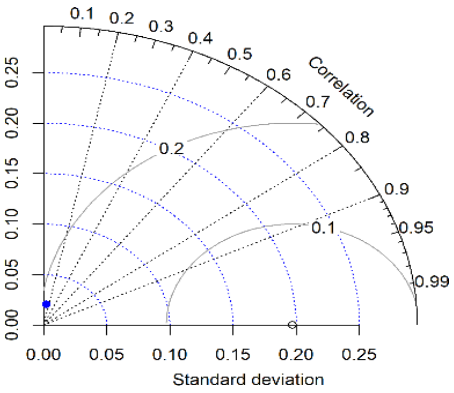


Figure 2(d). A Taylor diagram for a Random Forest model

Figure 2. A taylor diagram for different ML models

Table 2. OLS Results of a model summary for ANN.

Variable	*Estimate*	*Std_Error*	*t_value*	*p_value*	*CI_Lower*	*CI_Upper*
(Intercept)	98.77981	0.025771	3832.921	0	98.72902	98.83061
x	-0.07797	0.000755	-103.22	2.02E-186	-0.07946	-0.07648

Table 3. OLS results of a model summary for Decision Tree

Variable	*Estimate*	*Std_Error*	*t_value*	*p_value*	*CI_Lower*	*CI_Upper*
(Intercept)	98.9049	0.02504	3949.736	0	98.85557046	98.95427
x	-0.0817	0.00073	-111.4290	3.62E-194	-0.08323563	-0.080342

Table 4. OLS results of a model summary for Random Forest

Variable	*Estimate*	*Std_Error*	*t_value*	*p_value*	*CI_Lower*	*CI_Upper*
(Intercept)	98.89414	0.025091	3941.423	0	98.84469	98.94359
x	-0.08148	0.000735	-110.793	1.23E-193	-0.08293	-0.08004

Table 5. OLS results of a model summary for SVR

Variable	*Estimate*	*Std_Error*	*t_value*	*p_value*	*CI_Lower*	*CI_Upper*
(Intercept)	98.56865	0.032103	3070.401	0	98.50538	98.63193
x	-0.07918	0.000941	-84.143	3.48E-168	-0.08103	-0.07732

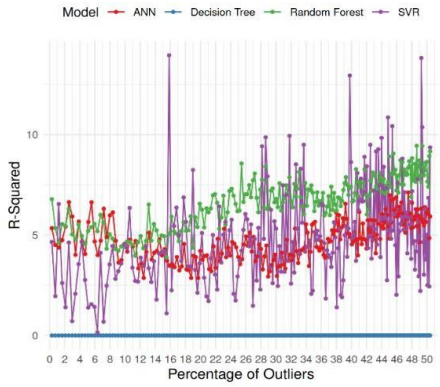


Figure 3. Percentage outiliers versus model fitting.

Trends of outlier effect on the model's training accuracy

Despite looking at the significance level of the model's accuracy in correlation and standard deviation between the observed and actual data, our study also considers training accuracy, error, and time taken to run a model in the presence of outliers. Figure 4(a) to Figure 4(d) presents the results showing how outliers affect the training accuracy of a machine learning model. Two aspects can be observed from the Figures. First, the model's initial parameters optimization causes slight upward trends, as observed at the beginning of Figure 4(a) to Figure 4(d). At this stage, the training accuracy is slightly improved due to the relatively low number of outlier definitions; as the outlier composition increases, the model's ability to generalize is impaired, and the accuracy subsequently declines. This phenomenon is frequently observed when early iterations are dominated by well-behaved data points, which causes an initial rise in performance before the anomalous observations become prevalent.

The second observation from Figure 4(a) to Figure 4(d) is that all four models are affected by the number of outliers in the training data. The accuracy decreases as the outlier percentage increases (from 0% to 50%)

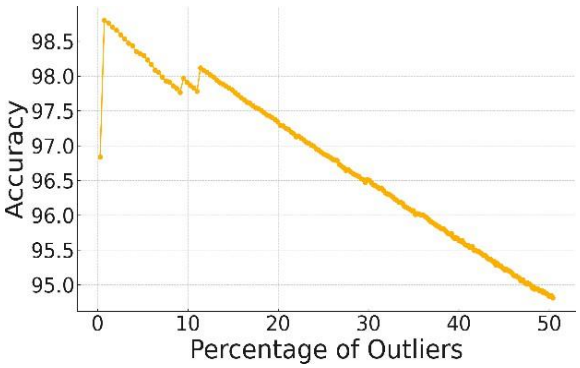


Figure 4(a). Outliers vs training accuracy for ANN.

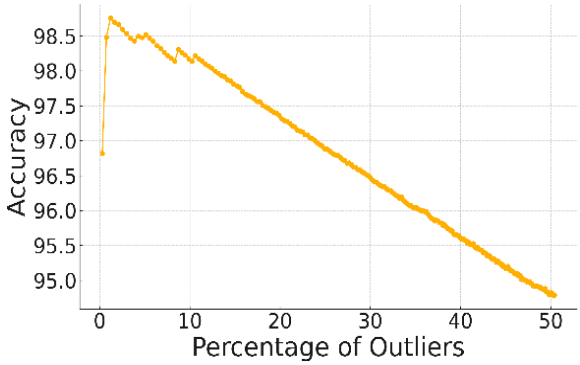


Figure 4(b). Outliers vs training accuracy for decision tree.

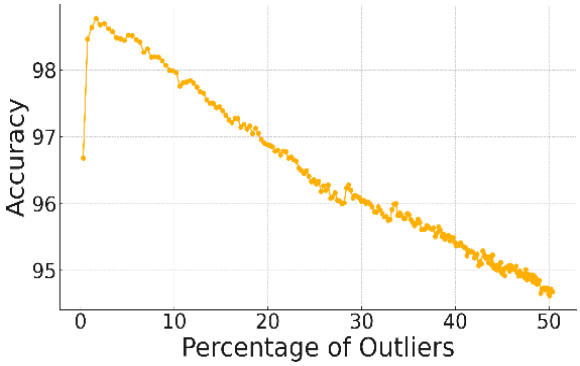


Figure 4(c). Outliers vs training accuracy for SVR.

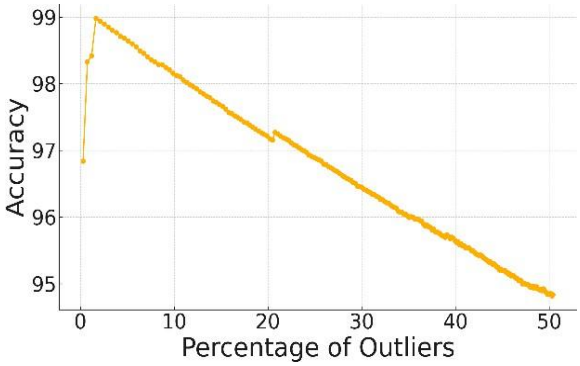


Figure 4(d). Outliers vs training accuracy for random forest.

Figure 4. Outliers vs training accuracy for different ML models.

The graphs show a zig-zag linear relationship (with a negative correlation) between the outlier percentage and the model's training accuracy.

Outlier effect on the model's training error

The study also compares the models under the specified number of outliers regarding MSE. First, the findings concerning the MSE values of ANN, SVR, Random Forest, and Decision Tree were presented. Figure 5 shows a combined graph of MSE versus percentage outliers. Figure 8 shows that the model's training error increases as the number of outliers increases.

As seen in the Figure 5, the error is less when no outliers exist in the training data, but it increases linearly as the percentage of outliers approaches 50%. Moreover, Figure 8 shows that the remaining three models have exhibited almost a similar trend, except for the SVR model, which produced higher MSE values.

Outlier effect on the model's training time

The study investigated how the model's training time was affected by the level of outliers in the data. The findings of the investigation are demonstrated through the graphs in Figure 6. The Figure shows that the training time of decision tree and random forest models is not explicitly affected by the increase in the number of outliers.

This observation contradicts the SVR and ANN models, where their training time increases slightly as the number of outliers increases in the data. However, the training time of the ANN model does not grow linearly, as observed with SVR. The relationship between percentage outliers and model training time in ANN is as exhibited in Figure 6

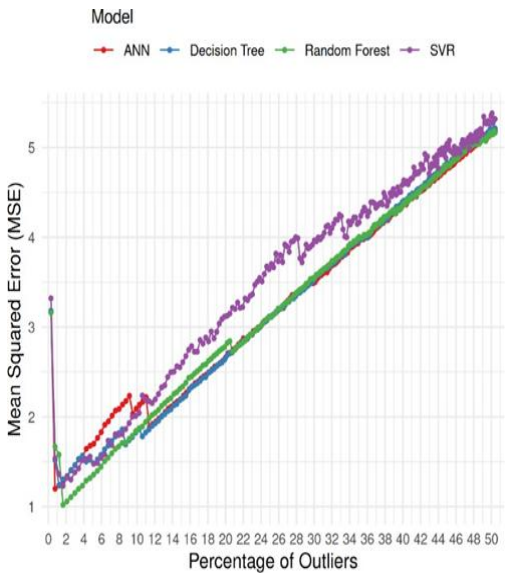


Figure 5. Percentage outliers versus MSE

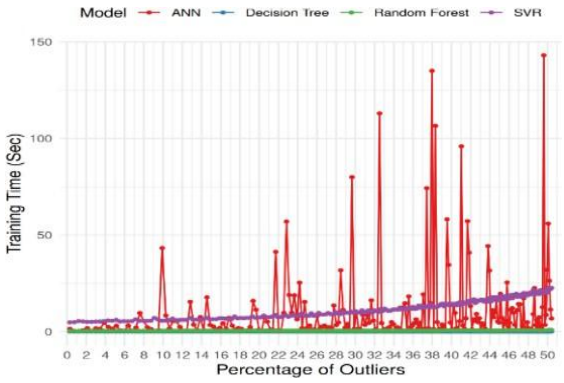


Figure 6. Percentage outliers versus training time.

5. Discussions

This study analyses the effect of outliers on a model's training accuracy, highlighting two major aspects: performance and accuracy. The analysis is based on a load dataset of short-term forecasting models using time of day and day-type variables. Four sample machine learning models, namely Artificial Neural Networks (ANN), Decision Tree, Random Forest, and Support Vector Regression (SVR), were used. The findings indicate that the model's performance and training accuracy vary significantly with the number of outliers in the data.

Effect of outliers on model training accuracy

The findings of this study reveal that the trend between outliers and model training accuracy is linear for SVR, DT, ANN, and RF. However, the trend of ANN and DT seems nonlinear when the outlier percentage is below a critical threshold of 13.978% (Figure 6 to Figure 9). In addition, the trends of SVR are observed to undergo a nonlinear trend with a zig-zag motion. This observation is also noticed in the case of model training error. The errors increase linearly as the number of outliers increases. The findings align with existing literature, indicating that SVR and DT follow a linear trend [32]. However, the findings reveal a nonlinear trend, particularly for ANN and DT, which contrasts with existing literature that suggests a linear trend across all models.

One possible reason behind these findings is that the impact of outliers on model performance may not be uniform; some outliers have a more substantial influence depending on their distance from the data distribution or their leverage within the feature space. This discrepancy could lead to abrupt drop in accuracy at specific thresholds (e.g., after 13.978%), disrupting a linear trend.

Additionally, the zig-zag motion may result from model sensitivity to slight variations in the outlier distribution, especially if the outliers are injected randomly or vary in magnitude. Our findings contrast with the existing literature [25], which assumes a linear degradation in accuracy with increasing outliers, possibly due to using synthetic or uniformly distributed noise. In contrast, our dataset may introduce more heterogeneity.

The observed trend between outliers and model accuracy or training errors is nonlinear, especially beyond the critical threshold of 13.978%

outliers, suggesting that outliers disproportionately affect model performance once their presence surpasses a certain point. The finding implies that machine learning models may need more sophisticated mechanisms to handle outliers, particularly when their frequency is high. Thisobservation calls for developing nonlinear models or more robust outlier-detection techniques that effectively handle the nonlinear effect on accuracy and errors.

Outliers and Model Performance

Results on performance show that the model's fitness decreases significantly as the number of outliers increases. In Du et al. [33], the results also show that the presence of outliers reduces the performance of the ML models after testing on three clusters of datasets. However, the authors tested only two ML models, whereas this study tested four, making our results more generalized.

In terms of the training time as an indicator of model performance, the time increases as the number of outliers increases. However, the findings reveal that the effect of outliers on the model's training time is not generic to all models [34]. For example, the training time of the Random Forest and Decision Tree models is significantly less affected by the number of outliers. In SVR, the training time increases gently as outliers increase. Regarding ANN, its training time increases nonlinearly while observing a random walk trend. Furthermore, the findings have revealed that the effect of outliers on the model's training performance is highly significant.

This significance underscores the complexity of training time and model performance dynamics. It suggests that models such as RF and DT are more robust in handling outliers without significant computational overhead. In contrast, other models (e.g., ANN and SVR) may require more efficient algorithms to handle training with high outlier presence.

In addition, model selection should consider accuracy, error rates, and training time, especially when dealing with outliers. If computational efficiency is a priority, Random Forest and Decision Trees may be more suitable. In contrast, SVR and ANN may require additional tuning to optimize performance in the presence of outliers without excessive computational burden.

The limitations of this study are twofold: First, the training data used in this study were based on a single source. Second, our experiments considered only four machine learning models (ANN, DT, RF, and SVR) as a case study, leaving other models unconsidered. These limitations might lead to some biases and ineffective generalization of the findings. However, the study has paved the way for researchers to address outlier data during the training of ML models.

6. Conclusion

Previous studies assume that outliers always cause a linear decline in model performance. This study offers new insights by showing the nonlinear relationship between outliers, model accuracy, and training errors. The study provides a general understanding of how ML models handle the presence of outliers. We have highlighted the effects of outliers on model performance and training accuracy, helping to guide experts and researchers on data preprocessing and model selection.

The findings of our study reveal that different models need distinct approaches to handle outliers, indicating that a one-size-fits-all strategy may be ineffective. Therefore, the authors emphasize the need for nonlinear models, such as stepwise and spline regression, an adaptive approach, and careful model selection when dealing with outliers. Given the new findings on the varying trends between outliers and model training accuracy, further research is needed to validate these results by examining other machine learning models and considering more data sources.

ACKNOWLEDGEMENT

The authors of this paper would like to sincerely thank the TANESCO team for their outstanding collaboration throughout the research, particularly in problem identification, data collection, and requirement elicitation.

CONTRIBUTIONS OF CO-AUTHORS

Hussein Bakiri	[ORCID: 0000-0003-0359-0117]	Conceived the idea and wrote the paper
Hadija Mbembati	[ORCID: 0000-0001-8748-4674]	Provided technical details and insights into the literature review and content writing

REFERENCES

[1] I. Price *et al.*, "Probabilistic weather forecasting with machine learning," *Nature*, vol. 637, no. April 2024, pp. 8–11, 2024, doi: 10.1038/s41586-024-08252-9.

[2] S. Patel and D. Kumar, "Predictive identification of oral cancer using AI and machine learning," *Oral Oncol. Reports*, vol. **13**, no. September 2024, p. 100697, 2025, doi: 10.1016/j.oor.2024.100697.

[3] J. X. Huang, X. Z. Shi, N. Zhang, Y. Q. Hu, and J. Q. Wang, "Prediction of bond strength between fibers and the matrix in UHPC utilizing machine learning and experimental data," *Mater. Today Commun.*, vol. **42**, p. 111136, Jan. 2025, doi: 10.1016/j.mtcomm.2024.111136.

[4] J. H. Sullivan, M. Warkentin, and L. Wallace, "So many ways for assessing outliers: What really works and does it matter?," *J. Bus. Res.*, vol. **132**, pp. 530–543, Aug. 2021, doi: 10.1016/j.jbusres.2021.03.066.

[5] J. W. Osborne and A. Overbay, "The power of outliers (and why researchers should ALWAYS check for them)," *Pract. Assessment, Res. Eval.*, vol. **9**, no. **6**, 2014.

[6] K. H. Lai, D. Zha, J. Xu, Y. Zhao, G. Wang, and X. Hu, "Revisiting Time Series Outlier Detection: Definitions and Benchmarks," *Thirty-fifth Conf. Neural Inf. Process. Syst. Datasets Benchmarks Track (Round1)*, no. NeurIPS, pp. 1–15, 2021, [Online]. Available: https://github.com/datamllab/tods/tree/benchmark.%0Ahttps://openreview.net/forum?id=r8IvOsn Hchr

[7] C. S. K. Dash, A. K. Behera, S. Dehuri, and A. Ghosh, "An outliers detection and elimination framework in classification task of data mining," *Decis. Anal. J.*, vol. **6**, no. January, p. 100164, 2023, doi: 10.1016/j.dajour.2023.100164.

[8] M. Al Samara, I. Bennis, A. Abouaissa, and P. Lorenz, "A Survey of Outlier Detection Techniques in IoT: Review and Classification," *J. Sens. Actuator Networks*, vol. **11**, no. **1**, 2022, doi: 10.3390/jsan11010004.

[9] W. Xu, J. Jang-Jaccard, A. Singh, Y. Wei, and F. Sabrina, "Improving Performance of Autoencoder-Based Network Anomaly Detection on NSL-KDD Dataset," *IEEE Access*, vol. **9**, pp. 140136–140146, 2021, doi: 10.1109/ACCESS.2021.3116612.

[10] D. Vasdekis, S. Yfantidou, S. Efstathiou, and A. Vakali, "WeMoD: A Machine Learning Approach for Wearable and Mobile Physical Activity Prediction," *2022 IEEE Int. Conf. Pervasive Comput. Commun. Work. other Affil. Events, PerCom Work. 2022*, pp. 385–390, 2022, doi: 10.1109/PerComWorkshops53856.2022.9767541.

[11] S. Saha, A. Haque, and G. Sidebottom, "Towards an Ensemble Regressor Model for Anomalous ISP Traffic Prediction," 2022, [Online]. Available: http://arxiv.org/abs/2205.01300

[12] M. Zanin, R. Lóczi, and A. Zanin, ""Outlier detection: Underused in sport science despite outliers' impact on inference and prediction"," *J. Sports Sci.*, vol. **00**, no. **00**, pp. 1–11, 2024, doi: 10.1080/02640414.2024.2443313.

[13] F. ODODO, N. A.-I. J. of African, and U. 2025, "UNDERSTANDING THE INFLUENCE OF OUTLIERS ON MACHINE LEARNING MODEL INTERPRETABILITY," *Int. J. AFRICAN*

Sustain. Dev. Res., vol. **7**, no. **2**, 2025, doi: 10.70382/tijasdr.v07i2.019.

[14] C. Li, D. Liu, M. Wang, H. Wang, and S. Xu, "Detection of Outliers in Time Series Power Data Based on Prediction Errors," *Energies*, vol. 16, no. 2, 2023, doi: 10.3390/en16020582.

[15] S. Mohammed *et al.*, "The effects of data quality on machine learning performance on tabular data," *Inf. Syst.*, vol. 132, pp. 1–50, 2025, doi: 10.1016/j.is.2025.102549.

[16] A. Kowalskie, "The Impact of Outliers on Linear Regression Models: Detection and Correction Strategies," *OTS Can. J.*, vol. **4**, no. **6**, pp. 108–118, 2025, doi: 10.58840/fzbcv732.

[17] S. Ansari *et al.*, "Impact of Outliers on Regression and Classification Models: An Empirical Analysis," in *Proceedings - International Conference on Developments in eSystems Engineering, DeSE*, IEEE, 2024, pp. 211–218. doi: 10.1109/DeSE63988.2024.10912020.

[18] U. Athiyah, S. Rustad, M. A. Soeleman, Muljono, and M. Akrom, "Effectiveness of K-means Clustering-Based Outlier Handling Techniques in Colonic Condition Endoscopy Image Classification Using CNN," in *COMNETSAT 2024 - IEEE International Conference on Communication, Networks and Satellite*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 276–282. doi: 10.1109/COMNETSAT63286.2024.10862507.

[19] A. Boublia *et al.*, "Enhancing precision in PANI/Gr nanocomposite design: Robust machine learning models, outlier resilience, and molecular input insights for superior electrical conductivity and gas sensing performance," *J. Mater. Chem. A*, vol. **12**, no. **4**, pp. 2209–2236, 2023, doi: 10.1039/d3ta06385b.

[20] F. A. Howes, "Introduction to Perturbation Techniques (Ali Hasan Nayfeh)," *SIAM Rev.*, vol. **24**, no. **3**, pp. 355–356, 1982, doi: 10.1137/1024080.

[21] T. J. III, S. M.-J. of D. Science, and U. 2025, "Impact of Data Perturbation for Statistical Disclosure Control on the Predictive Performance of Machine Learning Techniques," *Sch. Stat. Cent. Appl. Stat.*, pp. 1–20, 2025, Accessed: May 23, 2025. [Online]. Available: https://www.researchgate.net/profile/Sayed-Mostafa-4/publication/391100453_Impact_of_Data_Perturbation_for_Statistical_Disclosure_Control_on_the_Predictive_Performance_of_Machine_Learning_Techniques/links/680a594aded433155738f57 a/Impact-of-Data-Perturbati

[22] C. Mahlich, T. Vente, J. B.-I. C. On, and U. 2025, "From theory to practice: Implementing and evaluating e-fold cross-validation,"

[23] N. K. Elfaki *et al.*, "A Novel Deep Learning Approach for COVID-19 Case Identification: Enhanced CNN-KNN Hybrid Model with 5-Fold Cross-Validation," *Pakistan J. Life Soc. Sci.*, vol. **22**, no. **2**, pp. 723–748, 2024, doi: 10.57239/PJLSS-2024-22.2.0053.

[24] H. Azis, "Assessing the Performance of Logistic Regression in Heart Disease Detection through 5-Fold Cross-Validation," *Int. J. Artif. Intell. Med. Issues*, vol. **2**, no. **1**, pp. 1–11, 2024, doi: 10.56705/ijaimi.v2i1.137. B. Z.-E. of quality of life and well-being research and U. 2024, "Ordinary least-squares (OLS) model," *Springer*, pp. 4867–4869, 2024, doi: 10.1007/978-3-031-17299-1_2008.PDF.

[25] V. Magar, D. Ruikar, S. Bhoite, and R. Mente, "Innovative Inter Quartile Range-based Outlier Detection and Removal Technique for Teaching Staff Performance Feedback Analysis," *J. Eng. Educ. Transform.*,

vol. **37**, no. **3**, pp. 176–184, 2024, doi: 10.16920/jeet/2024/v37i3/24013.

[26] D. Dallah, H. Sulieman, and A. Alzaatreh, "Outlier detection based on range statistic empirical evaluation and comparisons," in *Statistical Outliers and Related Topics*, 2025, pp. 153–164. Accessed: May 23, 2025. [Online]. Available: https://books.google.com/books?hl=en&lr=&id=ty00EQAAQBAJ&oi=fnd&pg=PA153&dq=Interquartile+Range+(IQR)+for+Outlier+Detection%2Bdata+analysis&ots=OfE0b4tcOv&sig=Mv_HmHLEJXYflE2BiG1CdLJlJDM

[27] S. Gnat, "Impact of categorical variables encoding on property mass valuation," in *Procedia Computer Science*, Elsevier, Jan. 2021, pp. 3542–3550. doi: 10.1016/j.procs.2021.09.127.

[28] A. Dobbs, T. Elgindy, B.-M. Hodge, A. Florita, and J. Novacheck, "COMMON AND POPULAR MACHINE LEARNING ALGORITHMS," *ciir.in*, 2017, Accessed: Feb. 22, 2025. [Online]. Available: https://ciir.in/doc/publication/Intelligent Electrical Systems_978-81-962236-9-4.pdf#page=13

[29] H. Tissot, "FormulAI: Designing Rule-Based Datasets for Interpretable and Challenging Machine Learning Tasks," *Artif. Intell. Appl.*, vol. 3, no. 1, pp. 72–82, 2025, doi: 10.47852/bonviewAIA42021781.

[30] R. Al-Amri, R. K. Murugesan, M. Man, A. F. Abdulateef, M. A. Al-Sharafi, and A. A. Alkahtani, "A review of machine learning and deep learning techniques for anomaly detection in iot data," *Appl. Sci.*, vol. **11**, no. **12**, 2021, doi: 10.3390/app11125320.

[31] S. Karamcheti, R. Krishna, L. Fei-Fei, and C. D. Manning, "Mind your outliers! Investigating the negative impact of outliers on active learning for visual question answering," *ACL-IJCNLP 2021 - 59th Annu. Meet. Assoc. Comput. Linguist. 11th Int. Jt. Conf. Nat. Lang. Process. Proc. Conf.*, pp. 7265–7281, 2021, doi: 10.18653/v1/2021.acl-long.564.

[32] B. Du, P. D. Lund, and J. Wang, "Improving the accuracy of predicting the performance of solar collectors through clustering analysis with artificial neural network models," *Energy Reports*, vol. 8, pp. 3970–3981, 2022, doi: 10.1016/j.egyr.2022.03.013.

[33] M. Guimarães *et al.*, "Predicting Model Training Time to Optimize Distributed Machine Learning Applications," *Electron.*, vol. **12**, no. **4**, pp. 1–21, 2023, doi: 10.3390/electronics12040871.