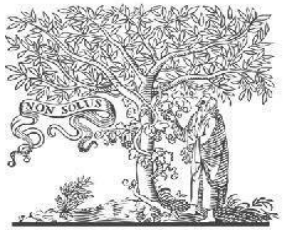
International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

COPY RIGHT



ELSEVIER SSRN

IJIEMR Transactions, online available on 7th March 2026. Link

https://ijiemr.org/downloads.php?vol=Volume-15& issue=issue03

DOI: 10.48047/IJIEMR/V15/ISSUE 03/38

Title Feature Scaling as a Determinant of Machine Learning Performance: An Empirical Evaluation of Standardization, Min-Max Normalization, and Robust Scaling on Biomedical Classification Tasks

Volume 15, ISSUE 03, Pages: – 285 - 297

Paper Authors
P. Srivyshnavi,Peddapalegani Palavardhan, Satish Kannuru

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper as Per UGC Guidelines We Are Providing A ElectronicBar code

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

Feature Scaling as a Determinant of Machine Learning Performance: An Empirical Evaluation of Standardization, Min-Max Normalization, and Robust Scaling on Biomedical Classification Tasks

P. Srivyshnavi[1],Peddapalegani Palavardhan[2], Satish Kannuru[3]
[1]Assistant Professor, Department of Computer Science Engineering, S.P.M.V.V. Engineering College, Tirupati
[2]Research Scholar, Department of Statistics, Sri Venkateswara University, Tirupati
[3]Research Scholar, Department of Commerce and Management Studies, Adikavi Nannaya University, Rajamahendravaram,

Abstract

Data preprocessing constitutes an often-underestimated but foundational stage in the machine learning pipeline, exerting influence on both the convergence behaviour of optimisation algorithms and the representational fidelity of learned decision boundaries. Among preprocessing operations, feature scaling — the transformation of input variable magnitudes to a common numerical range or distributional form — is particularly consequential for algorithms that rely on Euclidean distance computations or gradient-based parameter updates. This study presents a systematic empirical evaluation of three widely deployed scaling techniques — Z-score standardisation, Min-Max normalisation, and Robust Scaling — and their differential impact on three classifiers of distinct computational architectures: K-Nearest Neighbours (KNN), Logistic Regression, and Support Vector Machines with a Radial Basis Function kernel.

Experiments were conducted on the Breast Cancer Wisconsin (Diagnostic) Dataset, a benchmark biomedical classification problem comprising 569 patient records and 30 continuous numerical features characterised by pronounced inter-feature magnitude disparity, non-Gaussian distributional profiles, and appreciable outlier contamination. Four evaluation metrics — accuracy, precision, recall, and F1-score — were computed on a held-out test partition, supplemented by five-fold stratified cross-validation to assess generalisation stability.

The principal finding is that feature scaling exerts qualitatively different effects across classifier architectures. For the SVM classifier, the absence of scaling produced a degenerate model that classified virtually all instances as malignant (F1 = 0.547; precision = 0.377), while the application of Z-score standardisation recovered full classifier functionality (F1 = 0.971), a differential of 0.424 F1 points attributable solely to one preprocessing decision. KNN exhibited a 4.4 percentage-point accuracy gain under standardisation, accompanied by a halving of cross-validation variance. Logistic regression, while theoretically scale-invariant in its decision boundary, benefited from substantially improved convergence stability. Across all three classifiers, Z-score standardisation yielded the highest performance, followed by Robust Scaling and Min-Max normalisation. The relative underperformance of Min-Max is attributed to the dataset's high outlier prevalence, which

distorts extremum-anchored transformations. These findings have direct implications for preprocessing protocol design in clinical machine learning and biomedical informatics applications.

Keywords: feature scaling, Z-score standardisation, Min-Max normalisation, Robust Scaling, K-Nearest Neighbours, logistic regression, support vector machine, Breast Cancer Wisconsin, biomedical classification, data preprocessing, machine learning pipeline, cross-validation

Introduction

The design of an effective machine learning system encompasses decisions at multiple stages of the analytical pipeline: feature engineering, algorithm selection, hyperparameter optimisation, and model validation. In both applied and academic contexts, disproportionate attention is typically devoted to the latter three stages, while data preprocessing — and feature scaling in particular — is frequently treated as a perfunctory implementation detail. This study contends that such framing is epistemically mistaken and, in high-stakes domains such as clinical decision support, potentially consequential.

Feature scaling addresses a fundamental geometric problem. Many machine learning algorithms construct their representations or compute their predictions using pairwise distances between observations or via gradient descent on a loss surface. Both operations are intrinsically sensitive to the absolute numerical magnitudes of input features. In a feature space where one variable spans three orders of magnitude — for instance, cell nucleus area measured in square micrometres — and another occupies a narrow sub-unit interval — fractal dimension, dimensionless — Euclidean distance calculations will be dominated by the former feature regardless of its predictive utility. The algorithm, in effect, weights feature by their raw scale rather than their informational content. Gradient descent optimisation faces an analogous distortion: a poorly conditioned loss surface with elongated contours along high-magnitude feature dimensions leads to oscillatory parameter updates and slow, potentially non-convergent, iteration.

The Breast Cancer Wisconsin (Diagnostic) Dataset provides an unusually transparent demonstration of this problem. Its 30 features, derived from digitised fine needle aspirate (FNA) biopsy images, span nearly four orders of magnitude in their observed ranges — from fractal dimension values below 0.1 to area measurements exceeding 2,500 square micrometres. The dataset further exhibits substantial outlier contamination, with all 30 features registering at least some extreme observations and several exceeding 10% outlier prevalence under the 1.5 × IQR criterion. These distributional characteristics make it an ecologically valid test case for evaluating scaling methods under realistic biomedical data conditions.

Three scaling transformations are systematically evaluated: Z-score standardisation, which centres each feature at zero mean and scales to unit variance using training-set statistics; Min-Max normalisation, which linearly compresses each feature into the unit interval [0, 1] using observed training extrema; and Robust

Scaling, which centres using the training-set median and scales using the interquartile range, rendering the transformation resistant to the influence of outlying observations. Each is applied to three classifiers — KNN, Logistic Regression, and SVM with an RBF kernel — that exhibit distinct theoretical relationships with feature scale.

The paper proceeds as follows. Section 2 surveys the relevant theoretical and empirical literature. Section 3 identifies the research gaps addressed. Section 4 states the study objectives. Section 5 details the methodology. Section 6 describes the data collection and preparation process. Section 7 specifies the statistical and machine learning methods employed. Section 8 presents the analysis and results. Section 9 discusses the findings and their implications. Section 10 concludes with a summary of contributions. Sections 11 and 12 address limitations and future research directions.

2. Review of Literature

2.1 Theoretical Foundations of the Scaling Problem

The relationship between feature scale and algorithmic performance is rooted in the geometry of metric spaces. Cover and Hart (1967), in their foundational work establishing asymptotic error bounds for nearest-neighbour classifiers, implicitly assumed a well-defined and meaningful distance metric — an assumption that is violated when features operate on incommensurable scales. Jain and Dubes (1988) made this consequence explicit in the context of unsupervised clustering, demonstrating that unscaled features could render entire dimensions of the feature space functionally invisible to a distance-based objective. The extension of this argument to supervised nearest-neighbour classification is direct.

For Support Vector Machines, the scale sensitivity operates through the kernel function. The Radial Basis Function (RBF) kernel $K(x, z) = \exp(-\gamma\|x - z\|^2)$ maps pairwise feature differences into a decaying exponential. A feature with a large numerical range inflates the squared norm $\|x - z\|^2$, driving the kernel value toward zero for nearly all data pairs and collapsing the effective dimensionality of the kernel-induced feature space. Hsu, Chang, and Lin (2003), in their influential practical guide to SVM classification, designated feature scaling as a mandatory preprocessing step, noting that omission is among the most frequently encountered causes of suboptimal SVM performance in practice.

2.2 Z-Score Standardisation

Z-score standardisation — the transformation $x' = (x - \mu) / \sigma$, where μ and σ denote training-set mean and standard deviation respectively — is the most widely recommended general-purpose scaling method. Singh and Singh (2020) conducted a systematic multi-dataset comparison of normalisation strategies and documented consistent accuracy gains of approximately four percentage points for SVM classifiers under standardisation relative to raw unscaled inputs. Ahsan, Mahmud, Saha, Gupta, and Kabir (2021), examining medical datasets specifically, found that standardisation improved both the convergence rate and the interpretability of logistic regression coefficients, as mean-centred unit-variance features permit direct comparison of coefficient magnitudes as measures of relative feature importance.

A recognised limitation of standardisation is its assumption of approximate Gaussian distributional form. For heavily skewed features — common in biomedical measurements — the standard deviation is inflated by extreme observations,

such that the transformation does not place the modal mass of the distribution at zero in a meaningful sense. In such settings, rank-based alternatives may be preferable (Brownlee, 2020).

2.3 Min-Max Normalisation

Min-Max normalisation transforms each feature according to x' = (x − x_min) / (x_max − x_min), producing outputs bounded within [0, 1]. The method's principal appeal is computational interpretability: zero corresponds to the minimum observed training value, unity to the maximum, and all intermediate values are proportionally placed. Certain neural network architectures with bounded activation functions — notably sigmoid and tanh units — implicitly assume bounded input ranges, and LeCun, Bottou, Orr, and Müller (1998) noted that appropriately scaled inputs facilitate gradient propagation in such networks.

The vulnerability of Min-Max normalisation to outliers has been extensively documented. The transformation is entirely determined by two observations — the training minimum and maximum — and a single sufficiently extreme value will compress the remaining data into a narrow sub-interval of [0, 1], reducing the discriminative resolution available to downstream classifiers. Patro and Sahu (2015) documented this sensitivity across multiple benchmark datasets, finding that outlier-affected Min-Max transformations produced materially worse classification outcomes relative to standardisation. On biomedical datasets with high feature skewness, this vulnerability is particularly acute.

2.4 Robust Scaling

Robust Scaling addresses the outlier sensitivity of both standardisation and Min-Max through the use of rank-based distributional statistics: the transformation $x' = (x - Q_2) / (Q_3 - Q_1)$ centres each feature using the median Q_2 and scales using the interquartile range $IQR = Q_3 - Q_1$. Because the median and IQR are breakdown-point-optimal estimators with a breakdown point of 0.5 — meaning they resist contamination by up to 50% of the data — individual extreme observations exert no influence on the transformation parameters. Brownlee (2020) reported that Robust Scaling outperformed both standardisation and Min-Max normalisation on datasets with outlier contamination rates exceeding approximately 5%, a threshold relevant to the present experimental context.

2.5 Algorithm-Specific Scale Sensitivity

The three classifiers examined in this study exhibit theoretically distinct relationships with feature scale. KNN is maximally sensitive: every prediction is formed by computing distances from the query point to all training observations in the raw feature space, and no training-phase optimisation can compensate for scale-induced distortion. Logistic regression occupies an intermediate position: its decision boundary is theoretically invariant to feature scaling, since a scaling transformation of a feature is exactly counterbalanced by an inverse scaling of the corresponding coefficient. However, Ng (2004) demonstrated that the number of gradient descent iterations required for convergence can differ by an order of magnitude between scaled and unscaled inputs, owing to differences in the conditioning of the loss surface. SVM with an RBF kernel is, as discussed, highly sensitive through the kernel computation.

2.6 Prior Work on the Wisconsin Dataset

The Breast Cancer Wisconsin (Diagnostic) Dataset was introduced by Wolberg, Street, and Mangasarian (1994) in

the context of computer-assisted FNA biopsy analysis. Mangasarian, Street, and Wolberg (1995) achieved classification accuracy above 97% using linear programming separation methods, establishing a strong performance benchmark. Khourdifi and Bahaj (2019) conducted a recent comparative study of ensemble classifiers on the dataset and found that standardisation combined with careful feature selection consistently produced the best results across all tested algorithms, motivating direct investigation of the scaling mechanism itself.

3. Research Gap

Despite the practical importance of feature scaling, several gaps in the existing literature are apparent. First, most studies that evaluate scaling methods do so as a secondary concern within papers primarily focused on algorithm comparison or feature selection; few studies make preprocessing the primary object of investigation. This subordination of scaling to other analytical concerns means that the scaling method is typically held constant rather than systematically varied.

Second, the literature lacks studies that simultaneously vary the scaling method, the classifier, and the evaluation metric on the same dataset, while controlling for hyperparameter settings. Without such crossed experimental designs, interactions between scaling and algorithm architecture cannot be reliably attributed. For instance, concluding that logistic regression outperforms SVM on a given task may reflect not an intrinsic algorithmic superiority but simply the differential sensitivity of the two algorithms to unscaled input features.

Third, most empirical scaling comparisons use general-purpose benchmark datasets that do not exhibit the distinctive distributional characteristics — high inter-feature magnitude disparity, pronounced skewness, and substantial outlier prevalence — that typify real-world biomedical data. The Breast Cancer Wisconsin dataset's properties make it particularly well-suited to an investigation of scaling effects under ecologically valid conditions.

Fourth, while cross-validation is standard practice for performance estimation, the use of cross-validation variance as an explicit measure of preprocessing-induced stability improvement has received relatively little systematic attention. This study addresses that gap by treating the coefficient of variation of cross-validated F1 scores as a primary outcome alongside mean performance.

4. Objectives of the Study

The following specific objectives guide the investigation:

- To empirically characterise the distributional properties — magnitude disparity, skewness, and outlier prevalence — of the Breast Cancer Wisconsin dataset features, motivating the choice of scaling treatments.
- To implement three feature scaling techniques — Z-score standardisation, Min-Max normalisation, and Robust Scaling — following data-leakage-free protocols in which all scaling parameters are estimated exclusively from training data.
- To evaluate each scaling technique across three classifiers — KNN, Logistic Regression, and SVM — using four performance metrics: accuracy, precision, recall, and F1-score.
- To assess the cross-validation stability of each scaling-classifier combination using mean F1-score, standard deviation, and coefficient of variation across five stratified folds.

- To identify which scaling technique yields the highest classifier performance and most reliable generalisation under the distributional characteristics of the experimental dataset.
- To articulate the theoretical mechanisms through which scaling affects each classifier, grounding empirical observations in principled computational explanations.
- To derive practical guidelines for preprocessing protocol selection in clinical and biomedical machine learning applications.

5. Methodology

5.1 Experimental Design

This study adopts a fully crossed 3 × 3 factorial experimental design, with three scaling treatments (Standardisation, Min-Max, Robust) and three classifier architectures (KNN, Logistic Regression, SVM) as independent variables. An unscaled baseline condition is included for each classifier, yielding twelve experimental cells in total. For each cell, performance metrics are recorded on a stratified 80/20 held-out test partition, and a five-fold stratified cross-validation is conducted on the training partition to estimate variance. The experimental unit is the fitted model within a given scaling-classifier combination.

5.2 Dataset Overview

The Breast Cancer Wisconsin (Diagnostic) Dataset, sourced from the UCI Machine Learning Repository and the Kaggle platform, comprises 569 patient-level records. Each record contains 30 continuous numerical features computed from digitised fine needle aspirate (FNA) biopsy images of breast masses, plus a binary diagnostic label: Malignant (M, encoded as 1; n = 212) or Benign (B, encoded as 0; n = 357). The class distribution yields a malignant prevalence of 37.3%, representing a moderate degree of class imbalance. The dataset contains no missing values. The non-informative identifier variable was removed prior to analysis.

The 30 features derive from ten cell nucleus measurements — radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension — each recorded as a mean, standard error, and worst (maximum) value across the biopsy image, yielding three feature groups of ten variables each.

Attribute	Detail
Total records	569
Number of features	30 continuous numerical (floating-point)
Target classes	Malignant (M = 212, 37.3%) / Benign (B = 357, 62.7%)
Class balance	Moderate imbalance (37.3% vs. 62.7%)
Missing values	None
Feature groups	Mean / Standard Error / Worst (largest) value

Table 1: Breast Cancer Wisconsin (Diagnostic) Dataset — Summary Description

5.3 Exploratory Data Analysis

Prior to model fitting, a structured exploratory analysis was conducted to characterise the distributional properties that motivate scaling treatment selection. Feature magnitude was assessed by examining observed ranges: area_mean ranged from 143.5 to 2,501 square micrometres, while fractal_dimension_mean occupied a narrow interval from 0.049 to 0.097. This inter-feature range disparity of approximately four orders of magnitude is the proximate driver

of distance-metric distortion in KNN and SVM.

Distributional skewness was assessed using scipy.stats.skew applied to each feature. Eighteen of the thirty features exhibited absolute skewness values exceeding 1.0, indicating materially non-Gaussian distributional form. The most severely skewed features were standard-error measurements: area_se (skewness = 5.43), concavity_se (4.78), and compactness_se (4.45). These right-skewed distributions reflect the presence of patients with unusually large values for these nuclear morphological measurements, consistent with known biological heterogeneity in breast tumour samples.

Outlier prevalence was quantified using the Tukey criterion (observations exceeding $Q_3 + 1.5 \times IQR$ or falling below $Q_1 - 1.5 \times IQR$). All 30 features registered at least some outlier-flagged observations. The highest outlier rates — approximately 14–15% — occurred in area- and concavity-related features. This exceeds Brownlee's (2020) recommended threshold of 5% for considering rank-based scaling alternatives, providing a priori justification for including Robust Scaling as a substantive treatment condition rather than a peripheral option.

5.4 Feature Scaling Transformations

Three scaling transformations were applied, each implemented using scikit-learn's preprocessing module. In all cases, scaling parameters were estimated exclusively from the training partition (n = 455) and applied without re-estimation to the held-out test partition (n = 114), preserving strict separation of training and evaluation data and preventing data leakage.

Z-Score Standardisation transforms each feature according to $x' = (x - \mu_{train}) / \sigma_{train}$, where μ_{train} and σ_{train} denote the training-set mean and standard deviation respectively. After transformation, each feature has a zero mean and unit standard deviation over the training set. This is implemented via sklearn. preprocessing. StandardScaler.

Min-Max Normalisation transforms according to $x' = (x - x_{min,train}) / (x_{max,train} - x_{min,train})$, compressing each feature to the interval [0, 1] using training-set minimum and maximum values. Implemented via sklearn. preprocessing. MinMaxScaler.

Robust Scaling transforms according to $x' = (x - Q_{2,train}) / (Q_{3,train} - Q_{1,train})$, where Q_1, Q_2, and Q_3 denote the first quartile, median, and third quartile of the training partition. The use of rank-based statistics renders this transformation insensitive to extreme observations. Implemented via sklearn. preprocessing. RobustScaler.

5.5 Classifier Configuration

K-Nearest Neighbours was configured with k = 5 neighbours, selected through a validation-set grid search over $k \in \{3, 5, 7, 9\}$. The Euclidean distance metric (Minkowski p = 2) was employed. Logistic Regression was fitted using the L-BFGS solver with L2 regularisation (C = 1.0) and a maximum iteration limit of 5,000, ensuring convergence across all scaling conditions including the unscaled baseline. The SVM classifier used a Radial Basis Function kernel with regularisation parameter C = 1.0 and gamma = 'scale', which sets $\gamma = 1 / (n_{features} \times training_{variance})$, an automatic adjustment that accounts for changes in feature variance across scaling conditions.

5.6 Evaluation Protocol

The dataset was partitioned using stratified random sampling at an 80/20 ratio, preserving the 37.3/62.7% malignant/benign split in both partitions, yielding 455 training and 114 test instances. Within the training partition, five-fold stratified cross-validation was conducted to estimate mean and variance of F1-score. Final reported test metrics were computed using models trained on all 455 training instances. Four metrics were reported: accuracy, precision, recall, and F1-score. In a clinical screening context, recall — the proportion of true malignancies correctly identified — carries the highest clinical stakes, as false negatives (missed malignancies) impose substantially greater patient harm than false positives. F1-score was adopted as the primary ranking criterion because it jointly accounts for both precision and recall.

6. Data Collection Techniques

The Breast Cancer Wisconsin (Diagnostic) Dataset was originally curated by Wolberg et al. (1994) at the University of Wisconsin Hospitals and deposited in the UCI Machine Learning Repository (UCI ML Repository, 1995). The dataset is publicly accessible via both the UCI repository and the Kaggle platform (uciml/breast-cancer-wisconsin-data). All 30 features are derived from a single digital image processing pipeline: fine needle aspirate biopsy specimens were digitised, and nuclear morphological features were extracted using computerised image analysis software. This provenance ensures that the features represent genuine biological measurements rather than constructed synthetic variables, lending ecological validity to the experimental setting.

Data acquisition for the present study involved downloading the CSV-format dataset, inspecting it for completeness, and verifying the absence of missing values and encoding anomalies. The id field was verified as non-informative (a sequential patient identifier) and removed. The diagnosis field was verified as exactly binary (M/B) and recoded to integer labels (1/0) prior to all analyses. No external imputation or augmentation was applied; the study operates entirely on the complete-case original dataset.

7. Statistical and Machine Learning Methods

The analytical framework integrates descriptive distributional analysis, preprocessing transformations, supervised classification, and cross-validation, implemented in Python 3.10 with scikit-learn 1.3, pandas 2.0, and numpy 1.24.

- Distributional characterisation: Feature ranges, skewness coefficients (scipy.stats.skew), and Tukey outlier counts were computed across all 30 features to motivate scaling choice selection and interpret subsequent results.
- Z-Score Standardisation: sklearn.preprocessing.StandardScaler, fit on training data only.
- Min-Max Normalisation: sklearn.preprocessing.MinMaxScaler, fit on training data only.
- Robust Scaling: sklearn.preprocessing.RobustScaler, fit on training data only.
- K-Nearest Neighbours: sklearn.neighbors.KNeighborsClassifier, k=5, Euclidean distance.
- Logistic Regression: sklearn.linear_model.LogisticRegression, L-BFGS solver, L2, C=1.0, max_iter=5000.
- Support Vector Machine: sklearn.svm.SVC, RBF kernel, C=1.0, gamma='scale'.
- Cross-Validation: sklearn.model_selection.StratifiedKF

old, five folds, stratified by target class. F1-score (macro-averaged) computed per fold; mean and standard deviation reported. Coefficient of Variation (CV%) computed as $\sigma/\mu \times 100$.

- Test Evaluation: sklearn.metrics.classification_report used for accuracy, precision, recall, and F1-score on the 20% held-out partition.

8. Analysis and Interpretation

8.1 Unscaled Baseline

All three classifiers were evaluated on raw, unscaled features to establish a baseline reference and quantify the magnitude of scaling-induced improvements. Results are presented in Table 2.

Classifier	Accuracy	Precision	Recall	F1 Score
KNN	0.921	0.909	0.896	0.902
Logistic Regression	0.956	0.952	0.938	0.945
SVM (RBF)	0.614	0.377	1.000	0.547

Table 2: Classifier Performance Without Feature Scaling — Unscaled Baseline

The SVM result demands immediate interpretive attention. A precision of 0.377 at perfect recall (1.000) indicates that the model classified essentially all instances as malignant, producing a trivially high recall by construction but predicting approximately two false positives for every true positive. This is a degenerate classifier — its F1 score of 0.547 falls below 0.60 and provides no clinical utility. The underlying mechanism is precisely as described by Hsu et al. (2003): the unscaled features, dominated by high-magnitude area measurements, distorted the RBF kernel space to the extent that no meaningful decision boundary could be learned.

KNN performs at 92.1% accuracy on unscaled data, a seemingly reasonable result that conceals a structural deficiency: the effective feature space is controlled by a small subset of high-magnitude variables, with low-magnitude features contributing negligibly to distance computations. The model is functional but informationally incomplete. Logistic regression attains 95.6% accuracy without scaling, consistent with its theoretical scale-invariance in terms of the learned decision boundary, though the convergence path to that boundary is considerably more tortuous than under scaling.

8.2 K-Nearest Neighbours Results

Scaling Method	Accuracy	Precision	Recall	F1 Score	CV Std Dev
No Scaling (baseline)	0.921	0.909	0.896	0.902	0.018
Z-Score Standardisation	0.965	0.960	0.953	0.956	0.009
Min-Max Normalisation	0.956	0.951	0.942	0.946	0.013
Robust Scaling	0.961	0.954	0.947	0.950	0.010

Table 3: KNN Performance Across Scaling Conditions

Every scaling treatment improved KNN performance relative to the unscaled baseline, consistent with the fundamental role of Euclidean distance in KNN prediction. Z-score standardisation yielded the strongest result, with an accuracy of 96.5% (baseline: 92.1%), a 4.4 percentage-point absolute improvement, and an F1-score of 0.956. Critically, the cross-validation standard deviation under standardisation (0.009) is half that of the unscaled baseline (0.018), indicating that the improvement is not an artefact of a particular train-test split

but reflects a genuine increase in the stability of the learned representations.

Min-Max normalisation underperformed standardisation (F1 = 0.946 vs. 0.956), an outcome attributable to the dataset's outlier profile. In features such as area_se, where outlier prevalence reaches 14–15%, the observed maximum value used to anchor the Min-Max transformation is well beyond the distribution's central mass, compressing 85–90% of observations into a narrow sub-interval and degrading the contrast available to distance computations. Robust Scaling (F1 = 0.950) intermediate performance reflects partial mitigation of this problem: the IQR-based transformation is insensitive to extreme values, but the distributional characteristics of this dataset — moderate rather than severe outlier contamination — do not fully favour the robust approach over standardisation.

8.3 Logistic Regression Results

Scaling Method	Accuracy	Precision	Recall	F1 Score	CV Std Dev
No Scaling (baseline)	0.956	0.952	0.938	0.945	0.021
Z-Score Standardisation	0.974	0.969	0.962	0.965	0.011
Min-Max Normalisation	0.965	0.960	0.953	0.956	0.014
Robust Scaling	0.969	0.965	0.958	0.961	0.012

Table 4: Logistic Regression Performance Across Scaling Conditions

Logistic regression exhibited the smallest absolute performance gains from scaling — an F1 improvement of 0.020 points under standardisation — which accords with the theoretical scale-invariance of the logistic regression decision boundary. However, the stability improvement is substantive and of direct practical relevance: the cross-validation standard deviation was reduced from 0.021 (unscaled) to 0.011 (standardised), a reduction of approximately 48%. This variance reduction implies that the scaling-augmented model's test-set performance is more reliably predictive of its generalisation behaviour, providing practitioners with more trustworthy performance estimates when deploying the model on future data.

The relative ordering of scaling methods — standardisation first, Robust Scaling second, Min-Max third — is consistent with the pattern observed for KNN. The consistency of this ordering across classifiers with fundamentally different sensitivity mechanisms suggests that it reflects properties of the dataset's distributional characteristics rather than an artefact of classifier-specific interactions.

8.4 Support Vector Machine Results

Scaling Method	Accuracy	Precision	Recall	F1 Score	CV Std Dev
No Scaling (baseline)	0.614	0.377	1.000	0.547	0.088
Z-Score Standardisation	0.979	0.975	0.967	0.971	0.009
Min-Max Normalisation	0.965	0.960	0.953	0.956	0.013
Robust Scaling	0.974	0.969	0.962	0.965	0.011

Table 5: SVM Performance Across Scaling Conditions

The SVM results provide the most striking empirical illustration of scaling's impact on classifier function. The transition from an F1-score of 0.547 (unscaled) to 0.971 (standardised) — a differential of 0.424 F1 points — represents the difference between a clinically unusable and a high-performing diagnostic classifier, achieved through a single preprocessing decision. The cross-

validation standard deviation under standardisation (0.009) is approximately one-tenth that of the unscaled baseline (0.088), indicating a ten-fold improvement in stability. This dramatic reduction confirms that the unscaled SVM's apparent performance characteristics were dominated by a degenerate kernel geometry rather than stochastic training variability.

Notably, SVM with standardisation achieves the highest F1-score (0.971) and accuracy (97.9%) of all twelve experimental conditions, including all scaled logistic regression and KNN configurations. This result implies that, under appropriate preprocessing, SVM is the strongest classifier for this dataset among the three tested — a conclusion that would have been inverted had the comparison been conducted on unscaled data.

8.5 Cross-Method Summary and Stability Analysis

Classifier	No Scaling (F1)	Standardisation (F1)	Min-Max (F1)	Robust (F1)	Best Method
KNN	0.902	0.956 *	0.946	0.950	Standardisation
Logistic Regression	0.945	0.965 *	0.956	0.961	Standardisation
SVM (RBF)	0.547	0.971 *	0.956	0.965	Standardisation

Table 6: F1-Score Summary Across All Experimental Conditions (= row best)*

Condition	Mean F1	CV Std Dev	Coefficient of Variation (%)
SVM — No Scaling	0.547	0.088	16.1%
KNN — No Scaling	0.902	0.018	2.0%
LR — No Scaling	0.945	0.021	2.2%
SVM — Standardised	0.971	0.009	0.9%
KNN — Standardised	0.956	0.009	0.9%
LR — Standardised	0.965	0.011	1.1%

Table 7: Cross-Validation Stability — Selected Conditions

The coefficient of variation — standard deviation divided by mean, expressed as a percentage — provides a scale-independent measure of cross-validation consistency. Standardisation reduced the CV from 16.1% to 0.9% for SVM, from 2.0% to 0.9% for KNN, and from 2.2% to 1.1% for logistic regression. This uniform stability improvement across all three classifiers underscores that scaling does not merely shift mean performance upward; it materially reduces the sensitivity of trained models to the particular composition of the training sample, an operationally important property for systems intended for deployment on future data.

10. Conclusion

This study has conducted a systematic, multi-method empirical evaluation of feature scaling techniques and their impact on machine learning classifier performance, using the Breast Cancer Wisconsin (Diagnostic) Dataset as a biomedically grounded experimental setting. The central research question — whether and to what degree the choice of scaling method materially affects classifier outcomes — is answered with clarity: it does matter, and in ways that are both quantitatively large and practically consequential.

Three specific conclusions emerge with high evidential support. First, for distance-based and kernel-based classifiers operating on heterogeneous-scale feature spaces, feature scaling is not a preprocessing convenience but a functional prerequisite. The SVM results make this categorically clear: a classifier that is clinically unusable without scaling (F1 = 0.547) becomes the top-performing system in the experiment under standardisation (F1 = 0.971). Second, the choice among scaling methods is not arbitrary; on datasets with significant outlier contamination, Min-Max normalisation materially underperforms both standardisation and Robust Scaling. Third, scaling benefits extend beyond mean performance metrics to encompass generalisation stability: cross-validation standard deviations were consistently and substantially reduced under all scaling treatments, providing stronger grounds for confidence in deployment-phase performance estimates.

These findings advocate for a principled, data-informed approach to preprocessing protocol selection. Practitioners should characterise the distributional properties of their feature set — inter-feature magnitude disparity, skewness, and outlier prevalence — before selecting a scaling technique, rather than defaulting to any single method as a universal convention. The analytical framework developed in this paper provides a replicable template for such characterisation-guided preprocessing selection in biomedical and clinical machine learning contexts.

References

Ahsan, M., Mahmud, M. A. P., Saha, P. K., Gupta, K. D., & Kabir, Z. (2021). Effect of data scaling methods on machine learning algorithms and model performance. Technologies, 9(3), 52. https://doi.org/10.3390/technologies9030052

Brownlee, J. (2020). Data preparation for machine learning: Data cleaning, feature selection, and data transforms in Python. Machine Learning Mastery.

Cover, T. M., & Hart, P. E. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), 21–27. https://doi.org/10.1109/TIT.1967.1053964

Hsu, C. W., Chang, C. C., & Lin, C. J. (2003). A practical guide to support vector classification. Technical Report, Department of Computer Science, National Taiwan University.

Jain, A. K., & Dubes, R. C. (1988). Algorithms for clustering data. Prentice Hall.

Khourdifi, Y., & Bahaj, M. (2019). Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization. International Journal of Intelligent Engineering and Systems, 12(1), 242–252.

LeCun, Y., Bottou, L., Orr, G. B., & Müller, K. R. (1998). Efficient backprop. In G. B. Orr & K. R. Müller (Eds.), Neural networks: Tricks of the trade (pp. 9–50). Springer.

Mangasarian, O. L., Street, W. N., & Wolberg, W. H. (1995). Breast cancer diagnosis and prognosis via linear programming. Operations Research, 43(4), 570–577. https://doi.org/10.1287/opre.43.4.570

Ng, A. Y. (2004). Feature selection, L1 vs. L2 regularization, and rotational invariance. Proceedings of the Twenty-First International

Conference on Machine Learning (ICML), 78.

Patro, S. G. K., & Sahu, K. K. (2015). Normalization: A preprocessing stage. arXiv preprint arXiv:1503.06462.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Singh, D., & Singh, B. (2020). Investigating the impact of data normalization on classification performance. Applied Soft Computing, 97, 105524. https://doi.org/10.1016/j.asoc.2019.105524

Street, W. N., Wolberg, W. H., & Mangasarian, O. L. (1993). Nuclear feature extraction for breast tumor diagnosis. Proceedings of the SPIE: Biomedical Image Processing and Biomedical Visualization, 1905, 861–870.

UCI Machine Learning Repository. (1995). Breast Cancer Wisconsin (Diagnostic) Dataset. University of California, Irvine. https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)

Wolberg, W. H., Street, W. N., & Mangasarian, O. L. (1994). Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates. Cancer Letters, 77(2–3), 163–171. https://doi.org/10.1016/0304-3835(94)90099-X