International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

# Beyond Feature Importance: Investigating Predictive Sensitivity in Machine Learning Models

**Simran Sharma[1], Dr. Mahesh Mulani[2]**

[1]Visiting Faculty, Dept. of Computer Science, KSKV Kachchh University, Bhuj - Kachchh, India
[2]Head, Dept. of Computer Science, KSKV Kachchh University, Bhuj - Kachchh, India

**Abstract**
Feature importance is widely used in machine learning to assess the contribution of individual predictors to model performance and support the interpretation of model behaviour. However, it remains unclear whether features identified as highly important are also those to which a model is most sensitive when their values are disturbed. This study empirically investigates the relationship between feature importance and predictive sensitivity under controlled feature perturbations. Experiments are conducted on five classification datasets, namely Sonar, Ionosphere, Breast Cancer Wisconsin, Parkinson's Disease, and Spambase, using Logistic Regression, Random Forest, and Gradient Boosting to represent different modelling characteristics. Permutation importance is used to estimate model-specific feature importance. Individual features in the test data are then progressively perturbed using Gaussian noise at multiple intensity levels, while the trained models remain unchanged. Predictive sensitivity is measured through the resulting degradation in F1 score, and Spearman rank correlation is used to examine the correspondence between feature importance and feature-wise sensitivity. The results show that higher feature importance is associated with greater perturbation sensitivity in several dataset and model combinations. However, the relationship is not consistent across all experimental conditions or classifiers. Both the magnitude and statistical significance of the observed associations vary across datasets, models, and perturbation intensities, with stronger associations observed at higher perturbation levels in several cases. These findings suggest that feature importance and perturbation sensitivity capture related but distinct aspects of model behaviour. Therefore, feature importance alone may not fully reflect how strongly model performance responds to disturbances in individual features. Controlled feature perturbation analysis can thus provide a complementary perspective for interpreting feature relevance and evaluating the robustness of machine learning models.

**Keywords:** Feature Importance, Predictive Sensitivity, Model Robustness, Permutation Importance, Machine Learning

## 1. Introduction

Machine learning models now support prediction and decision making in many application areas. As their use has expanded, understanding how individual input features contribute to their predictions has become an important part of model interpretation. Feature importance methods provide a way to estimate the relative influence of features on model predictions. However, a feature identified as highly

important may not necessarily be the feature to which a model is most sensitive when its values are disturbed.

This distinction becomes especially relevant when input data are imperfect, since measurement errors, noise, and other variations can alter feature values. Previous studies have shown that modifying influential features can affect predictive performance and that the resulting effect may vary across machine learning models [4]. Feature perturbation has also been used to evaluate the reliability of feature importance estimates by observing changes in model performance after selected features are modified [5]. Furthermore, perturbation-based approaches have demonstrated a relationship between feature relevance and changes in model predictions [6]. These studies establish an important connection between feature importance and model response to altered inputs.

Both permutation importance and Gaussian perturbation sensitivity quantify the degradation in predictive performance that results from corrupting a feature, the former through shuffling, the latter through controlled noise. Given this shared basis, some degree of correspondence between the two measures would be expected. What remains unclear is whether this correspondence is consistent across complete feature rankings, different classifiers, and progressively increasing perturbation levels, or whether it varies in ways that limit the reliability of permutation importance as a proxy for a feature's true sensitivity. This question therefore requires further empirical investigation.

The present study examines this relationship using five classification datasets, namely Sonar, Ionosphere, Breast Cancer Wisconsin, Parkinson's Disease, and Spambase. Logistic Regression, Random Forest, and Gradient Boosting are employed to represent different modelling approaches. Feature relevance is estimated using permutation importance, while predictive sensitivity is measured by introducing controlled Gaussian perturbations to individual features and observing the resulting change in F1 score. Spearman rank correlation is subsequently used to determine the correspondence between feature importance and perturbation sensitivity across datasets, classifiers, and perturbation intensities.

The study specifically investigates whether highly important features tend to exhibit greater predictive sensitivity, whether this relationship changes with increasing perturbation intensity, and whether similar patterns are observed across different datasets and classifiers. Through this analysis, the study aims to provide a clearer empirical understanding of the relationship between feature importance and predictive sensitivity and to determine whether the two measures provide consistent or complementary information about model behaviour.

## 2. Literature Review

Feature importance has become an important component of machine learning interpretation because it provides information about the relative contribution of input variables to model predictions. However, the reliability of an importance estimate depends on the method used, the characteristics of the data, and the behaviour of the underlying predictive model. Recent research has therefore moved beyond the estimation of feature importance alone and has examined the reliability and consistency of importance estimates under different modelling and data conditions. Antoska Knights et al. compared feature importance across different machine learning models and interpretability approaches [1]. Claborne et al. investigated the consistency of feature attribution across deep learning architectures [2], while Lee et al. examined the validity of feature importance under variations in predictive performance [3]. Collectively, these studies indicate that feature importance should be interpreted in relation to the model and conditions under which it is estimated.

Perturbation based analysis provides another approach for examining feature relevance and model behaviour. Padró Ferragut et al. proposed a model independent framework in which permutation importance was used to identify the most influential attribute and controlled noise was subsequently introduced into that attribute [4]. Their experiments demonstrated that perturbing a critical attribute can produce substantial degradation in predictive performance, although the extent of degradation varies across datasets and learning algorithms. Brocki and Chung examined feature perturbation as a mechanism for evaluating importance estimators and observed that a reduction in predictive performance after perturbing important features can provide evidence regarding the quality of an explanation [5]. They also noted that perturbation may introduce distributional artifacts, which should be considered when interpreting changes in model performance.

A related direction has examined feature importance through the magnitude of input modifications required to alter model predictions. Chapman Rounds et al. introduced Feature Importance by Minimal Adversarial Perturbation, which combines feature importance with counterfactual explanations by identifying perturbations capable of changing model classifications [6]. This work demonstrates that model response to changes in input features can provide information that complements conventional importance scores. Abdlatef et al. further examined the relationship between feature importance similarity and adversarial transferability in explainable intrusion detection systems [7]. Although their study addresses an adversarial setting, it further illustrates the connection between feature relevance and model response to changes in input features.

The effect of noisy or corrupted information on model behaviour has also been investigated directly. Moldovanu et al. evaluated machine learning and neural network classifiers under different forms of corrupted data, including feature noise and label noise [8]. Their findings showed that classifiers exhibit different levels of robustness and sensitivity when the quality of input information deteriorates. Pau et al. examined the robustness of feature selection methods under class noisy data and demonstrated that the stability of selected features can vary across feature selection methods and noise conditions [9]. These findings indicate that the effect of data disturbance depends on the characteristics of the data and the analytical method being used.

Recent studies have further questioned whether feature importance values should be treated as stable descriptions of model behaviour. Hwang et al. demonstrated that feature-based explanations generated using SHAP can change substantially under different feature representations, even when commonly used data engineering transformations are applied [10]. This suggests that an importance value is dependent on the representation and evaluation conditions under which it is obtained. From a broader perspective, Freiesleben and Grote examined robustness in machine learning in relation to stability under specified changes or interventions [11]. Together, these studies reinforce the importance of considering model behaviour under changing input or evaluation conditions rather than interpreting feature importance values in isolation.

Collectively, the existing literature establishes connections among feature importance, input perturbation, and predictive robustness. Previous studies have investigated the perturbation of influential attributes, the use of perturbation for evaluating importance estimates, classifier behaviour under noisy data, and the stability of feature-based explanations [4, 5, 8, 10]. However, comparatively less attention has been given to determining whether the complete ranking of feature importance corresponds to the complete ranking of predictive sensitivity when every feature is independently subjected to progressively increasing controlled perturbations. The present study addresses this aspect by examining

the rank-based association between permutation importance and F1 score degradation across multiple datasets, classifiers, and perturbation intensities. This enables feature relevance and predictive sensitivity to be evaluated as related but distinct characteristics of model behaviour.

## 3. Research Methodology

The study was designed to empirically examine whether the importance assigned to an individual feature corresponds to the sensitivity of a machine learning model when that feature is progressively perturbed. A common experimental procedure was applied across all datasets and classifiers to maintain consistency in feature importance estimation, perturbation, and performance evaluation. The methodology consisted of dataset preparation, model training, permutation importance estimation, controlled feature perturbation, sensitivity measurement, and statistical analysis.

### 3.1.Datasets

The experiments used five publicly available classification datasets that differ in their sample sizes, number of predictors, and overall data characteristics. The datasets used were Sonar, Ionosphere, Breast Cancer Wisconsin Diagnostic, Parkinson's Disease, and Spambase [12, 13, 14, 15, 16]. All five datasets represent binary classification problems and contain predominantly numerical input features, making them suitable for applying a consistent feature perturbation procedure.

**Table 1: Characteristics of the Selected Datasets**

| Dataset | Samples | Features | Classes |
|---|---|---|---|
| Sonar | 208 | 60 | 2 |
| Ionosphere | 351 | 34 | 2 |
| Breast Cancer Wisconsin Diagnostic | 569 | 30 | 2 |
| Parkinson's Disease | 195 | 22 | 2 |
| Spambase | 4,601 | 57 | 2 |

The selected datasets also differ considerably in dimensionality and sample size. Sonar contains 60 input features, Ionosphere contains 34 features, Breast Cancer Wisconsin Diagnostic contains 30 features, Parkinson's Disease contains 22 predictive features, and Spambase contains 57 features. This variation enables the relationship between feature importance and predictive sensitivity to be examined under different data conditions rather than relying on a single dataset.

### 3.2.Machine Learning Models

Three supervised classifiers were considered: Logistic Regression, Random Forest, and Gradient Boosting. Their inclusion provides three different modelling structures: linear classification, bagged decision trees, and sequentially boosted trees.

An 80:20 stratified train-test split was applied to each dataset, preserving the class distribution in both partitions. Model training was performed only on the training data, while baseline evaluation and all subsequent perturbation experiments were conducted on the test data. Standardization was performed using parameters estimated from the training data, and the same transformation was applied to the test data. The standardized data were used consistently across all three classifiers. The same train and test partition was retained throughout the experiment to ensure comparability between baseline and perturbed performance.

### 3.3.Feature Importance Estimation

Permutation importance was used to estimate the relevance of individual input features. This approach measures the change in predictive performance produced when the values of a particular feature are randomly permuted while the remaining features are left unchanged [17]. A greater reduction in performance indicates greater dependence of the trained model on that feature.

Permutation importance was calculated separately for every dataset and classifier using the F1 score as the evaluation measure. This produced an importance value and corresponding ranking for each feature. The resulting rankings were subsequently compared with the sensitivity rankings obtained through controlled feature perturbation.

### 3.4.Controlled Feature Perturbation

Predictive sensitivity was evaluated by independently perturbing one feature at a time in the test dataset. Gaussian noise was introduced according to the standard deviation of the feature being examined. For feature j, the perturbed value was defined as:

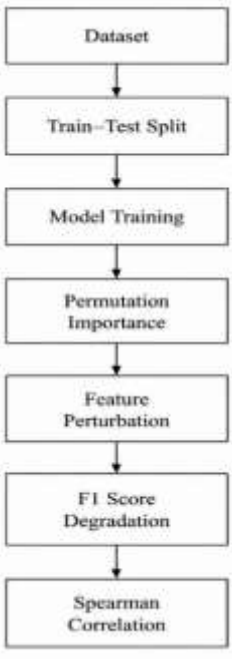
$$X_j + \lambda\sigma_j Z, \qquad Z \sim N(0,1)$$

where $X_j$ represents the original standardized feature value, $\sigma_j$represents the standard deviation of the corresponding standardized feature estimated from the training data, $Z$ is a random variable sampled from a standard normal distribution, and $\lambda$controls the intensity of perturbation.

Five nonzero perturbation levels were considered: 0.1, 0.2, 0.3, 0.4, and 0.5 times the standard deviation of the corresponding feature. The unmodified test data served as the baseline condition, corresponding to a perturbation level of 0.0. Only one feature was modified during each experiment, while all other feature values remained unchanged. The training data were not perturbed, since the analysis evaluated the response of models trained on clean data to feature disturbances introduced at test time.

To reduce the influence of random variation introduced by Gaussian noise, each perturbation experiment was repeated 10 times. The resulting F1 scores were averaged for each feature, model, dataset, and perturbation level.

**Figure 1. Experimental Framework**



### 3.5.Predictive Sensitivity Measurement

The sensitivity of a feature was quantified using the change in F1 score between the clean test data and the perturbed test data. The performance degradation was calculated as:

$$V_{j,\lambda} = \frac{1}{10}\sum_{r=1}^{10}\left(F1_{\text{clean}} - F1_{j,\lambda,r}\right)$$

Where $F1_{\text{clean}}$represents the F1 score on the unmodified test data, $F1_{j,\lambda,r}$ represents the F1 score obtained after perturbing feature $j$ at intensity $\lambda$ during repetition $r$, and $V_{j,\lambda}$represents the mean predictive sensitivity of feature $j$ at that perturbation level. A larger positive value indicates greater predictive sensitivity. Negative values were retained because they indicate cases in which the perturbed data produced a higher F1 score than the clean baseline.

To obtain an overall feature level sensitivity measure, predictive sensitivity was averaged across the five nonzero perturbation levels:

$$V_j = \frac{1}{5}\sum_{\lambda\in\{0.1,0.2,0.3,0.4,0.5\}} V_{j,\lambda}$$

Where $V_j$ represents the overall predictive sensitivity of feature $j$. This measure summarizes the average performance degradation associated with each feature across the five perturbation intensities.

### 3.6.Statistical Analysis

Spearman rank correlation was used to examine the association between permutation importance and predictive sensitivity. The Spearman coefficient was selected because the objective was to determine whether features receiving higher importance rankings also tended to exhibit greater performance degradation, without requiring a linear relationship between the two measures.

Correlation analysis was performed separately for each dataset, classifier, and nonzero perturbation level. An additional overall correlation was calculated between feature importance and the mean perturbation sensitivity of each feature. The correlation coefficient and corresponding p value were used to examine the direction, magnitude, and statistical significance of the observed associations. Because both permutation importance and predictive sensitivity were evaluated using changes in F1 score, some degree of correspondence between the two measures may be expected. The correlation analysis was therefore used to examine the extent to which their feature rankings correspond across different datasets, classifiers, and perturbation intensities, rather than to treat the two measures as theoretically independent.

In addition to correlation analysis, features were ranked according to permutation importance and the sensitivity patterns of the five highest ranked and five lowest ranked features were compared across increasing perturbation levels. Feature correlation matrices were also examined as a diagnostic measure because strong relationships among predictors can influence the interpretation of permutation-based importance values.

The complete experimental procedure therefore enables feature importance and predictive sensitivity to be evaluated independently and subsequently compared under consistent conditions across multiple datasets, classifiers, and perturbation intensities.

All experiments were conducted using a fixed random state of 42 for reproducibility where applicable. Logistic Regression was configured with a maximum of 5,000 iterations, Random Forest used 300 estimators, and Gradient Boosting was applied using its standard implementation settings. Permutation

importance was estimated using 10 repetitions, and each Gaussian perturbation experiment was independently repeated 10 times.

## 4. Results and Analysis

The experimental results were analyzed to determine whether features receiving higher permutation importance also exhibited greater predictive sensitivity under controlled perturbation. The analysis considered baseline model performance, changes in F1 score under increasing perturbation intensity, the association between feature importance and sensitivity, and differences across datasets and classifiers.

### 4.1.Baseline Model Performance

Baseline performance was first evaluated on the unmodified test data for Logistic Regression, Random Forest, and Gradient Boosting. Accuracy, F1 score, and ROC AUC were used to establish the predictive performance of each trained model before introducing feature perturbations. These baseline results served as the reference for calculating subsequent changes in F1 score.

**Table 2: Baseline Classification Performance**

| Dataset | Model | Accuracy | F1 | ROC_AUC |
|---|---|---|---|---|
| Sonar | Logistic Regression | 0.833 | 0.844 | 0.905 |
| Sonar | Random Forest | 0.833 | 0.844 | 0.924 |
| Sonar | Gradient Boosting | 0.857 | 0.864 | 0.925 |
| Ionosphere | Logistic Regression | 0.930 | 0.947 | 0.924 |
| Ionosphere | Random Forest | 0.958 | 0.967 | 0.972 |
| Ionosphere | Gradient Boosting | 0.972 | 0.978 | 0.971 |
| Breast Cancer Wisconsin | Logistic Regression | 0.982 | 0.986 | 0.995 |
| Breast Cancer Wisconsin | Random Forest | 0.947 | 0.958 | 0.994 |
| Breast Cancer Wisconsin | Gradient Boosting | 0.956 | 0.966 | 0.991 |
| Parkinsons | Logistic Regression | 0.923 | 0.949 | 0.924 |
| Parkinsons | Random Forest | 0.923 | 0.949 | 0.971 |
| Parkinsons | Gradient Boosting | 0.923 | 0.947 | 0.969 |
| Spambase | Logistic Regression | 0.929 | 0.909 | 0.970 |
| Spambase | Random Forest | 0.946 | 0.930 | 0.983 |
| Spambase | Gradient Boosting | 0.939 | 0.922 | 0.983 |

The models demonstrated different levels of predictive performance across the five datasets, reflecting differences in dataset characteristics and modelling approaches. The purpose of this comparison was not to identify a superior classifier, but to establish the clean performance of each model against which perturbation induced changes could be measured.

### 4.2.Feature Importance and Perturbation Sensitivity

Permutation importance produced a feature ranking independently for each dataset and classifier. Controlled Gaussian perturbations were then applied to individual features at progressively increasing intensity levels. In general, perturbing highly ranked features resulted in larger changes in F1 score than

perturbing features with low permutation importance. However, this pattern was not observed uniformly across all datasets and classifiers.

The comparison between the five highest ranked and five lowest ranked features provided a clear representation of this behaviour. For several dataset and model combinations, the mean F1 score degradation of the highly important features increased as perturbation intensity increased, whereas the degradation associated with the lower ranked features remained comparatively small. This distinction was particularly visible for Spambase and for several models evaluated on the Breast Cancer Wisconsin Diagnostic and Parkinson's Disease datasets.

At the same time, some experiments produced small or negative changes in F1 score. A negative value indicates that the perturbed test data produced a slightly higher F1 score than the clean test data for that experimental condition. Such values were retained in the analysis rather than being converted to zero, since they represent the observed response of the trained model to the applied perturbation.

**4.3.Association Between Feature Importance and Predictive Sensitivity**

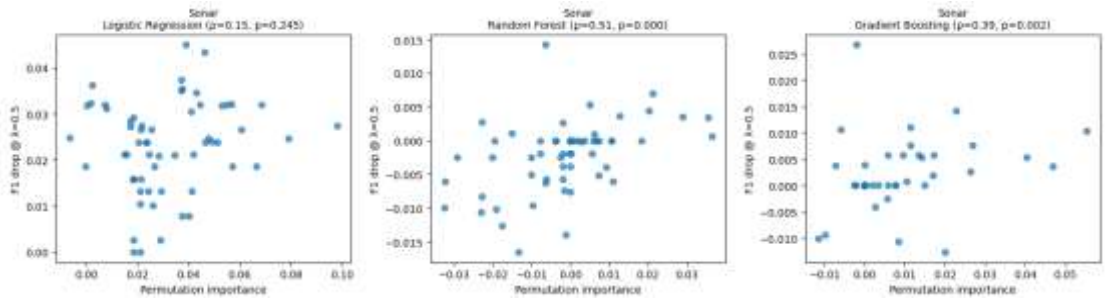
Spearman rank correlation was used to quantify the correspondence between permutation importance and mean perturbation sensitivity. Positive rank associations appeared in many dataset-classifier combinations, although their strength and statistical significance were not uniform.

The Breast Cancer Wisconsin Diagnostic dataset demonstrated a comparatively consistent positive association across all three classifiers. The overall Spearman coefficients were approximately 0.78 for Logistic Regression, 0.71 for Random Forest, and 0.74 for Gradient Boosting, with statistically significant results. Parkinson's Disease also exhibited positive associations across all three models, with coefficients of approximately 0.61, 0.65, and 0.59 for Logistic Regression, Random Forest, and Gradient Boosting respectively.

For Spambase, a positive association was also observed across the three classifiers. Logistic Regression produced a comparatively high correlation of approximately 0.76, while Random Forest and Gradient Boosting produced coefficients of approximately 0.46 and 0.50 respectively. These results indicate that features ranked as more important generally tended to exhibit greater sensitivity to perturbation, although the strength of this correspondence depended on the classifier.

Ionosphere showed stronger associations for Logistic Regression and Random Forest, with overall coefficients of approximately 0.65 and 0.73 respectively, while Gradient Boosting produced a weaker overall association of approximately 0.32. Sonar presented a different pattern. Logistic Regression and Random Forest showed relatively weak overall associations of approximately 0.20, whereas Gradient Boosting produced a higher correlation of approximately 0.40. The observed variation shows that the correspondence between importance and sensitivity depends on the dataset and modelling condition.

**Figure 2. Relationship Between Feature Importance and Predictive Sensitivity at 0.5 Standard Deviation**



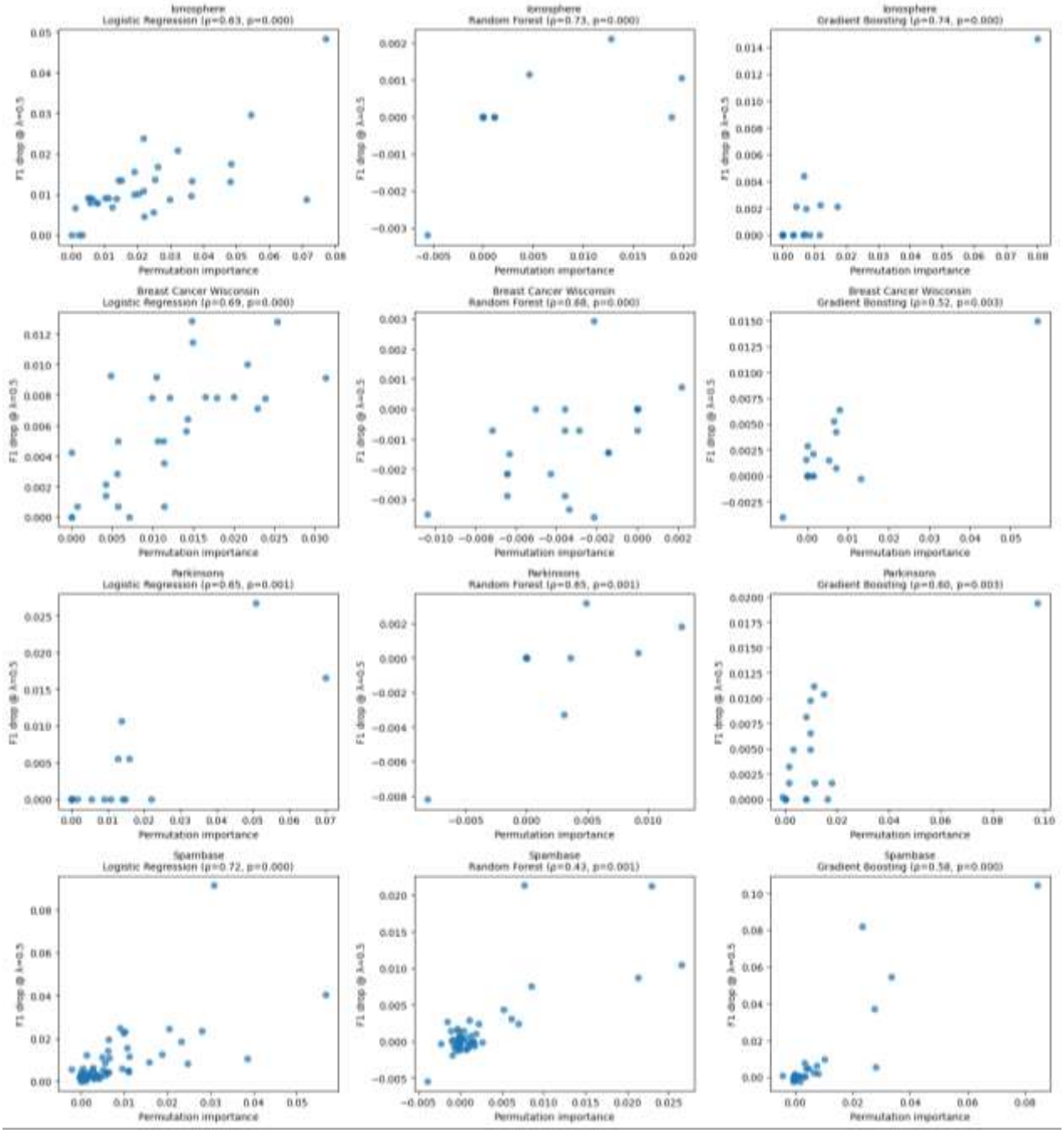



**Table 3: Spearman Correlation Between Feature Importance and Predictive Sensitivity Across Perturbation Levels**

| Dataset | Model | Corruption level sigma | Overall, Spearman ρ | p values |
|---|---|---|---|---|
| **Sonar** | Logistic Regression | overall | 0.200 | 0.125 |
| **Sonar** | Logistic Regression | 0.1 | 0.168 | 0.199 |
| **Sonar** | Logistic Regression | 0.2 | 0.119 | 0.367 |

| | | | | |
|---|---|---|---|---|
| **Sonar** | Logistic Regression | 0.3 | 0.299 | 0.020 |
| **Sonar** | Logistic Regression | 0.4 | 0.177 | 0.177 |
| **Sonar** | Logistic Regression | 0.5 | 0.153 | 0.245 |
| **Sonar** | Random Forest | overall | 0.201 | 0.124 |
| **Sonar** | Random Forest | 0.1 | -0.217 | 0.095 |
| **Sonar** | Random Forest | 0.2 | 0.108 | 0.410 |
| **Sonar** | Random Forest | 0.3 | 0.033 | 0.803 |
| **Sonar** | Random Forest | 0.4 | 0.172 | 0.189 |
| **Sonar** | Random Forest | 0.5 | 0.510 | 3.16E-05 |
| **Sonar** | Gradient Boosting | overall | 0.400 | 0.002 |
| **Sonar** | Gradient Boosting | 0.1 | 0.102 | 0.440 |
| **Sonar** | Gradient Boosting | 0.2 | 0.244 | 0.060 |
| **Sonar** | Gradient Boosting | 0.3 | 0.238 | 0.067 |
| **Sonar** | Gradient Boosting | 0.4 | 0.412 | 0.001 |
| **Sonar** | Gradient Boosting | 0.5 | 0.391 | 0.002 |
| **Ionosphere** | Logistic Regression | overall | 0.649 | 3.30E-05 |
| **Ionosphere** | Logistic Regression | 0.1 | 0.627 | 7.30E-05 |
| **Ionosphere** | Logistic Regression | 0.2 | 0.528 | 0.001 |
| **Ionosphere** | Logistic Regression | 0.3 | 0.403 | 0.018 |
| **Ionosphere** | Logistic Regression | 0.4 | 0.580 | 3.24E-04 |
| **Ionosphere** | Logistic Regression | 0.5 | 0.635 | 5.52E-05 |
| **Ionosphere** | Random Forest | overall | 0.726 | 1.19E-06 |
| **Ionosphere** | Random Forest | 0.1 | 0.394 | 0.021 |
| **Ionosphere** | Random Forest | 0.2 | 0.394 | 0.021 |
| **Ionosphere** | Random Forest | 0.3 | 0.394 | 0.021 |
| **Ionosphere** | Random Forest | 0.4 | 0.394 | 0.021 |
| **Ionosphere** | Random Forest | 0.5 | 0.726 | 1.19E-06 |
| **Ionosphere** | Gradient Boosting | overall | 0.319 | 0.066 |
| **Ionosphere** | Gradient Boosting | 0.1 | 0.216 | 0.221 |
| **Ionosphere** | Gradient Boosting | 0.2 | 0.043 | 0.807 |

| | | | | |
|---|---|---|---|---|
| **Ionosphere** | Gradient Boosting | 0.3 | 0.130 | 0.465 |
| **Ionosphere** | Gradient Boosting | 0.4 | 0.453 | 0.007 |
| **Ionosphere** | Gradient Boosting | 0.5 | 0.745 | 4.42E-07 |
| **Breast Cancer Wisconsin** | Logistic Regression | overall | 0.784 | 2.93E-07 |
| **Breast Cancer Wisconsin** | Logistic Regression | 0.1 | 0.414 | 0.023 |
| **Breast Cancer Wisconsin** | Logistic Regression | 0.2 | 0.680 | 3.56E-05 |
| **Breast Cancer Wisconsin** | Logistic Regression | 0.3 | 0.755 | 1.46E-06 |
| **Breast Cancer Wisconsin** | Logistic Regression | 0.4 | 0.766 | 8.00E-07 |
| **Breast Cancer Wisconsin** | Logistic Regression | 0.5 | 0.687 | 2.76E-05 |
| **Breast Cancer Wisconsin** | Random Forest | overall | 0.712 | 1.03E-05 |
| **Breast Cancer Wisconsin** | Random Forest | 0.1 | 0.480 | 0.007 |
| **Breast Cancer Wisconsin** | Random Forest | 0.2 | 0.347 | 0.060 |
| **Breast Cancer Wisconsin** | Random Forest | 0.3 | 0.685 | 3.00E-05 |
| **Breast Cancer Wisconsin** | Random Forest | 0.4 | 0.694 | 2.13E-05 |
| **Breast Cancer Wisconsin** | Random Forest | 0.5 | 0.682 | 3.26E-05 |
| **Breast Cancer Wisconsin** | Gradient Boosting | overall | 0.737 | 3.47E-06 |
| **Breast Cancer Wisconsin** | Gradient Boosting | 0.1 | 0.266 | 0.156 |
| **Breast Cancer Wisconsin** | Gradient Boosting | 0.2 | 0.793 | 1.70E-07 |
| **Breast Cancer Wisconsin** | Gradient Boosting | 0.3 | 0.822 | 2.64E-08 |
| **Breast Cancer Wisconsin** | Gradient Boosting | 0.4 | 0.705 | 1.35E-05 |
| **Breast Cancer Wisconsin** | Gradient Boosting | 0.5 | 0.516 | 0.003 |
| **Parkinsons** | Logistic Regression | overall | 0.612 | 0.002 |

| | | | | |
|---|---|---|---|---|
| **Parkinsons** | Logistic Regression | 0.1 | — | — |
| **Parkinsons** | Logistic Regression | 0.2 | — | — |
| **Parkinsons** | Logistic Regression | 0.3 | 0.712 | 2.03E-04 |
| **Parkinsons** | Logistic Regression | 0.4 | 0.553 | 0.008 |
| **Parkinsons** | Logistic Regression | 0.5 | 0.652 | 0.001 |
| **Parkinsons** | Random Forest | overall | 0.653 | 9.75E-04 |
| **Parkinsons** | Random Forest | 0.1 | 0.636 | 0.001 |
| **Parkinsons** | Random Forest | 0.2 | 0.761 | 3.99E-05 |
| **Parkinsons** | Random Forest | 0.3 | 0.810 | 4.84E-06 |
| **Parkinsons** | Random Forest | 0.4 | 0.303 | 0.171 |
| **Parkinsons** | Random Forest | 0.5 | 0.648 | 0.001 |
| **Parkinsons** | Gradient Boosting | overall | 0.594 | 0.004 |
| **Parkinsons** | Gradient Boosting | 0.1 | 0.243 | 0.275 |
| **Parkinsons** | Gradient Boosting | 0.2 | 0.191 | 0.395 |
| **Parkinsons** | Gradient Boosting | 0.3 | 0.167 | 0.457 |
| **Parkinsons** | Gradient Boosting | 0.4 | 0.290 | 0.190 |
| **Parkinsons** | Gradient Boosting | 0.5 | 0.604 | 0.003 |
| **Spambase** | Logistic Regression | overall | 0.763 | 4.99E-12 |
| **Spambase** | Logistic Regression | 0.1 | 0.749 | 2.12E-11 |
| **Spambase** | Logistic Regression | 0.2 | 0.745 | 3.07E-11 |
| **Spambase** | Logistic Regression | 0.3 | 0.743 | 3.59E-11 |
| **Spambase** | Logistic Regression | 0.4 | 0.768 | 3.02E-12 |
| **Spambase** | Logistic Regression | 0.5 | 0.715 | 4.09E-10 |
| **Spambase** | Random Forest | overall | 0.456 | 3.64E-04 |
| **Spambase** | Random Forest | 0.1 | 0.307 | 0.020 |
| **Spambase** | Random Forest | 0.2 | 0.406 | 0.002 |
| **Spambase** | Random Forest | 0.3 | 0.447 | 4.91E- |

| | | | | |
|---|---|---|---|---|
| | | | | 04 |
| **Spambase** | Random Forest | 0.4 | 0.397 | 0.002 |
| **Spambase** | Random Forest | 0.5 | 0.430 | 8.36E-04 |
| **Spambase** | Gradient Boosting | overall | 0.499 | 7.90E-05 |
| **Spambase** | Gradient Boosting | 0.1 | 0.356 | 0.007 |
| **Spambase** | Gradient Boosting | 0.2 | 0.313 | 0.018 |
| **Spambase** | Gradient Boosting | 0.3 | 0.461 | 3.07E-04 |
| **Spambase** | Gradient Boosting | 0.4 | 0.435 | 7.30E-04 |
| **Spambase** | Gradient Boosting | 0.5 | 0.581 | 2.14E-06 |

At low perturbation intensities, several datasets–model combinations exhibited identical or undefined rank correlations. For the Ionosphere dataset with Random Forest, the Spearman coefficient remained unchanged ($\rho = 0.39$) across perturbation levels of 0.1 to 0.4 standard deviations, since predictive performance was unaffected by noise for the majority of features at these intensities; only one feature showed a measurable change, producing an identical rank structure across these levels. For the Parkinson's Disease dataset with Logistic Regression, no feature exhibited any change in F1 score at perturbation levels of 0.1 and 0.2 standard deviations, resulting in a constant sensitivity vector for which the Spearman correlation is mathematically undefined; these cells are reported as "—" in Table 3. Both patterns reflect genuine model robustness to small perturbations on comparatively small test sets, rather than a computational error.

**4.4.Effect of Perturbation Intensity**

The relationship between feature importance and predictive sensitivity also changed with perturbation intensity. Small perturbations often produced little change in F1 score. Consequently, some experiments showed weak or statistically insignificant rank associations at the lower perturbation levels. As the magnitude of perturbation increased, clearer differences among features emerged in several dataset and classifier combinations.

This behaviour was particularly evident for Gradient Boosting on the Ionosphere dataset. The association between importance and sensitivity was weak at lower perturbation levels but increased considerably at higher levels. At a perturbation intensity of 0.4 standard deviations, the Spearman coefficient reached approximately 0.45, while at 0.5 standard deviations it increased to approximately 0.75. This result suggests that the relationship between feature importance and predictive sensitivity may become more apparent when the applied perturbation is sufficiently large to produce measurable differences in model performance.

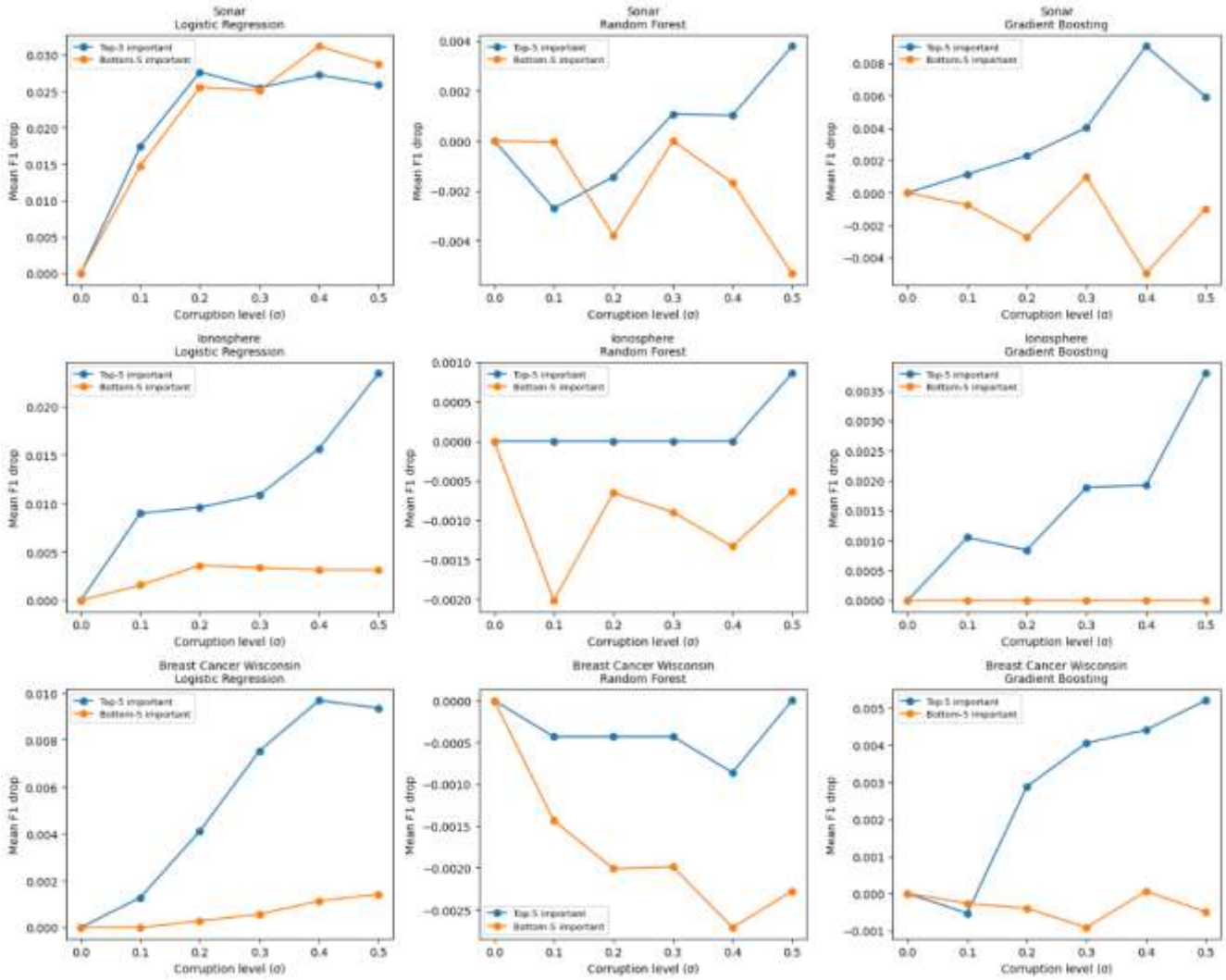
The effect was not identical across all datasets and models, indicating that perturbation intensity interacts with both the characteristics of the input data and the behaviour of the trained classifier. Consequently, evaluating sensitivity at a single perturbation level may provide an incomplete representation of the relationship between feature importance and model response.

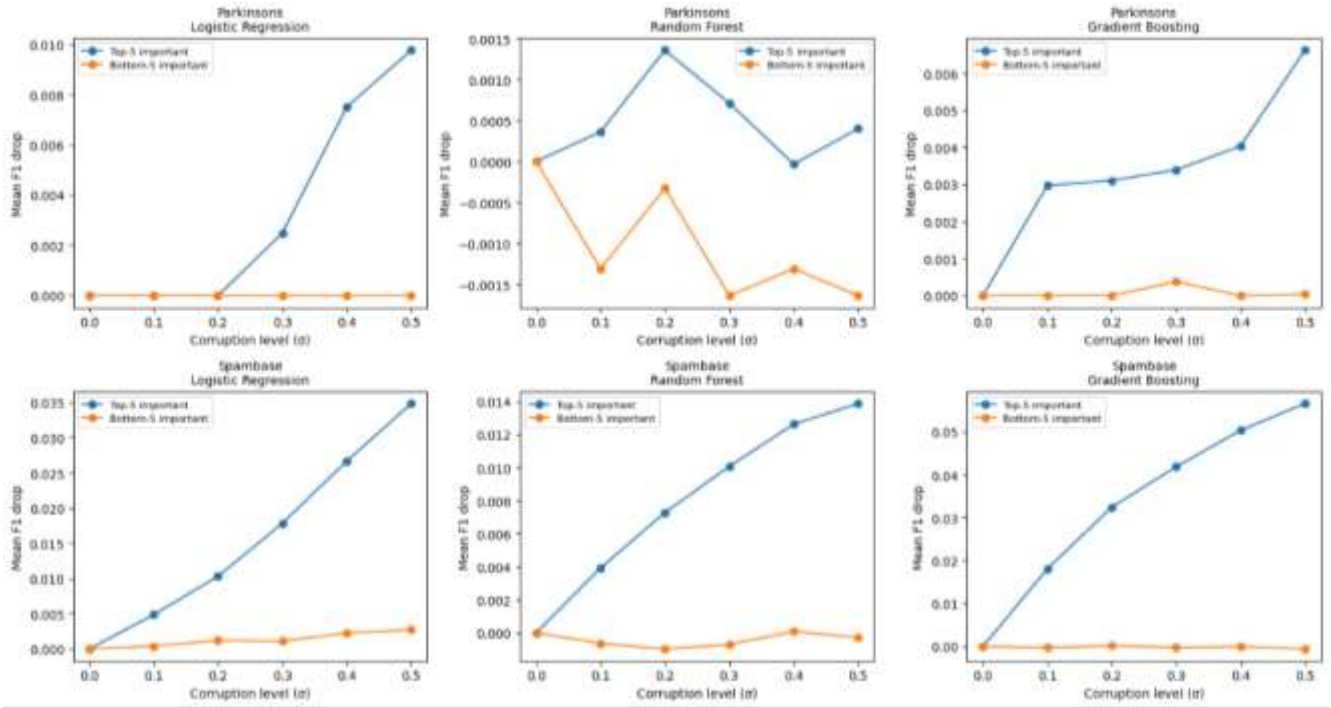
### 4.5.Overall Findings

The experimental results provide evidence of a positive association between feature importance and predictive sensitivity in a substantial number of the evaluated conditions. Features with higher permutation importance frequently produced greater F1 score degradation when perturbed. Nevertheless, the relationship was not universal. Its magnitude varied across datasets, classifiers, and perturbation intensities.

These findings indicate that permutation importance and perturbation sensitivity capture related aspects of model behaviour, but they should not be considered equivalent measures. A high importance ranking may indicate greater sensitivity in many cases, but the strength of this correspondence depends on the experimental context. The results therefore support the use of controlled perturbation analysis as a complementary procedure when interpreting feature importance and examining the predictive robustness of machine learning models.

**Figure 3 Predictive Sensitivity of Top Five and Bottom Five Important Features**






## 5. Discussion

Some positive correspondence between the two measures was expected because both are based on changes in predictive performance after feature information is disrupted. The experimental results confirm this association in several dataset and classifier combinations, but the strength of this relationship varies considerably — and in some conditions weakens or disappears entirely. Features assigned higher permutation importance generally produced larger reductions in F1 score when perturbed, particularly for the Breast Cancer, Parkinson's Disease, and Spambase datasets. However, this pattern was not consistently observed across all models and datasets. The weaker associations obtained for Sonar and some Ionosphere models demonstrate that feature importance cannot be considered a direct measure of perturbation sensitivity. This finding is consistent with previous work showing that changes to influential features can affect predictive performance, while the magnitude of the effect varies across datasets and learning algorithms [4, 5].

Perturbation intensity also affected the relationship observed between the two measures. At lower perturbation levels, changes in feature values often produced only small variations in predictive performance, resulting in weak or statistically insignificant rank correlations. As perturbation intensity increased, stronger associations emerged in several experiments. This behaviour was particularly evident for Ionosphere with Gradient Boosting, where the relationship became substantially stronger at higher perturbation levels. This suggests that the correspondence between feature importance and predictive sensitivity may become more apparent when the disturbance applied to a feature is sufficiently large to influence model predictions. Previous perturbation-based studies have similarly demonstrated that model responses depend on the nature and magnitude of modifications applied to input features [4, 6]. The stronger associations observed at higher perturbation intensities may partly reflect the increasing degree of feature disruption. Since permutation importance also evaluates performance after disrupting the information carried by a feature, stronger Gaussian perturbations may produce model responses that

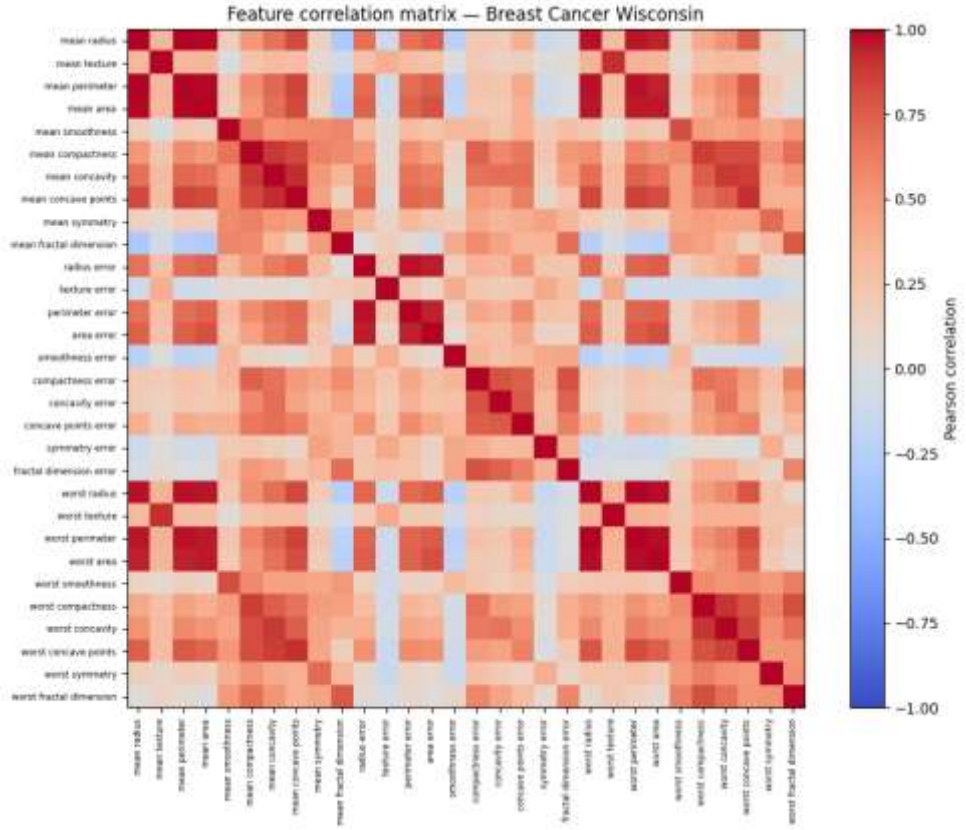
more closely resemble those induced by permutation. This provides one possible explanation for the increased correspondence observed at higher perturbation levels.

The relationship also differed by classifier. Even when evaluated on the same dataset, Logistic Regression, Random Forest, and Gradient Boosting did not always produce comparable importance-sensitivity patterns. Such variation is expected because these models represent different approaches to learning relationships between predictors and the target variable. Previous studies have also reported variation in feature importance across machine learning models and interpretability approaches [1]. Consequently, the effect of perturbing an individual feature depends not only on its estimated importance but also on how the trained model uses that feature together with other predictors.

The comparison between the five most important and five least important features provides additional support for this interpretation. In several experiments, particularly for Spambase and Breast Cancer, perturbation of highly ranked features resulted in greater average F1 score degradation than perturbation of features with low importance. Nevertheless, this separation was not equally clear in every experiment. Small negative F1 score changes were also observed in some cases, indicating that individual perturbations can occasionally improve predictive performance rather than reduce it. These variations further support treating feature importance and predictive sensitivity as related but distinct measures rather than assuming that one directly determines the other.

Correlation and redundancy among predictors may provide another explanation for differences between the two rankings. When multiple features contain similar predictive information, perturbing one feature may have a limited effect because information from related features remains available to the model. Similarly, permutation importance may distribute or underestimate the apparent contribution of correlated predictors. Previous research has demonstrated that feature importance and attribution can vary according to data representation, modelling conditions, and characteristics of the input features [2, 3, 10]. Therefore, disagreement between importance and sensitivity rankings does not necessarily indicate an unreliable model. Instead, it may reflect interactions among features, redundancy in the data, and differences in the way individual classifiers use the available information. The correlation structure illustrated in Figure 4 shows relationships among several predictors in the Breast Cancer Wisconsin Diagnostic dataset. Such relationships may influence the interpretation of permutation importance because predictive information associated with one feature may also be represented by correlated predictors.

**Figure 4. Feature Correlation Matrix for the Breast Cancer Wisconsin Diagnostic Dataset**




Overall, the findings suggest that feature importance and perturbation sensitivity provide related but distinct information about model behaviour. Permutation importance describes the contribution of a feature to predictive performance, whereas controlled perturbation analysis evaluates how model performance responds when the values of that feature are progressively disturbed. This interpretation is consistent with the broader view of robustness as the stability of model behaviour under specified changes or interventions [11]. Considering both measures can therefore provide a more complete interpretation of feature relevance and model robustness than relying on feature importance alone.

## 6. Limitations

Several aspects of the experimental design limit the scope of these findings. First, the experiments were conducted on five binary classification datasets containing predominantly numerical features. Although these datasets provide variation in sample size and dimensionality, the findings may not directly generalize to multiclass problems, categorical data, or substantially higher dimensional datasets.

Predictive sensitivity was examined only through Gaussian perturbation. Although this allowed feature values to be modified systematically across increasing noise levels, real-world disturbances may follow different patterns.

Third, permutation importance was used as the common feature importance method across all classifiers. The behaviour of other interpretation methods may differ, particularly when features are strongly correlated or contain redundant predictive information. Consequently, the observed relationship between importance and sensitivity should be interpreted specifically in the context of permutation importance.

Predictive sensitivity was evaluated using changes in F1 score. Since F1 is calculated from discrete class predictions, small changes in model outputs that do not alter the predicted class may not be reflected in

this measure. Future studies could complement F1 based sensitivity with probability-based measures such as log loss or changes in predicted probabilities.

Each experiment perturbed a single feature while leaving all other predictors unchanged. This experimental design enables the effect of individual features to be examined systematically, but it does not capture the combined effect of simultaneous perturbations across multiple features.

Finally, the analysis is limited to Logistic Regression, Random Forest, and Gradient Boosting. These models provide different learning characteristics, but additional algorithms may exhibit different relationships between feature importance and predictive sensitivity. Future research may therefore extend the analysis to other feature importance methods, perturbation strategies, datasets, and machine learning models. Additionally, at low perturbation intensities some model–dataset combinations produced constant or near-constant sensitivity vectors, limiting the interpretability of rank correlation in those specific conditions.

## 7. Conclusion

This study examined the relationship between feature importance and predictive sensitivity under controlled feature perturbations across five classification datasets and three machine learning models. Permutation importance was used to estimate feature relevance, while predictive sensitivity was measured through changes in F1 score after progressively introducing Gaussian perturbations to individual features.

Features ranked as more important were often more sensitive to perturbation, although this pattern did not hold consistently across every dataset, classifier, or perturbation level. Stronger associations were observed in several experimental settings, while weaker relationships in others demonstrate that a high importance value does not necessarily imply an equally high sensitivity to feature perturbation. The results also indicate that the strength of this relationship can change as perturbation intensity increases.

Overall, feature importance and predictive sensitivity appear to describe related aspects of model behaviour, but the experimental results do not support treating them as interchangeable measures. Although some agreement between them may be expected, this relationship is not guaranteed to remain consistent in practice. Feature importance provides information about the relevance of a predictor to model performance, whereas perturbation analysis provides additional information about how strongly predictive performance responds when that predictor is disturbed. Therefore, considering both perspectives can provide a more comprehensive understanding of model behaviour than relying on feature importance alone.

Future research can extend this analysis by considering additional feature importance techniques, alternative forms of perturbation, multiclass datasets, and a wider range of machine learning models. The influence of feature interactions and correlated predictors on the relationship between importance and sensitivity also provides an important direction for further investigation.

## References

1. Antoska Knights V., Mazlami V., Prchkovska M., Gajdoš Kljusurić J., “A Unified Interpretability Framework for Feature Importance in Machine Learning Models”, Algorithms, 2026, 19 (7), 548. https://doi.org/10.3390/a19070548
2. Claborne D., Flores J., Erwin S., Durell L., Degnan D., Richardson R., Fore R., Bramer L., “Consistency of Feature Attribution in Deep Learning Architectures for Multi-Omics”, Scientific

Reports, 2026, 16, 25890. https://doi.org/10.1038/s41598-026-58312-5
3. Lee Y., Baruzzo G., Kim J., Seo J., Di Camillo B., "Validity of Feature Importance in Low-Performing Machine Learning for Tabular Biomedical Data", IEEE Access, 2025, 13, 176381–176391. https://doi.org/10.1109/ACCESS.2025.3618851
4. Padró-Ferragut C., Martínez-Plumed F., Ramírez-Quintana M.J., "Robustness under Noise: Assessing the Impact of Perturbed Key Attributes on Machine Learning Models", International Journal of Data Science and Analytics, 2026, 22, Article 71. https://doi.org/10.1007/s41060-025-01006-4
5. Brocki L., Chung N.C., "Feature Perturbation Augmentation for Reliable Evaluation of Importance Estimators in Neural Networks", Pattern Recognition Letters, 2023, 176, 131–139. https://doi.org/10.1016/j.patrec.2023.10.012
6. Chapman-Rounds M., Bhatt U., Pazos E., Schulz M.A., Georgatzis K., "FIMAP: Feature Importance by Minimal Adversarial Perturbation", Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35 (13), 11433–11441. https://doi.org/10.1609/aaai.v35i13.17362
7. Abdlatef A., Abdulkareem M., Cakir D., "Estimating Adversarial Transferability from Feature-Importance Similarity in Explainable Intrusion Detection Systems", International Journal of Information Security, 2026, 25, Article 159. https://doi.org/10.1007/s10207-026-01333-y
8. Moldovanu S., Munteanu D., Sîrbu C., "Impact on Classification Process Generated by Corrupted Features", Big Data and Cognitive Computing, 2025, 9 (2), 45. https://doi.org/10.3390/bdcc9020045
9. Pau S., Perniciano A., Pes B., Rubattu D., "An Evaluation of Feature Selection Robustness on Class Noisy Data", Information, 2023, 14 (8), 438. https://doi.org/10.3390/info14080438
10. Hwang H., Bell A., Fonseca J., Pliatsika V., Stoyanovich J., Whang S.E., "SHAP-Based Explanations Are Sensitive to Feature Representation", Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, 2025, 1588–1601. https://doi.org/10.1145/3715275.3732105
11. Freiesleben T., Grote T., "Beyond Generalization: A Theory of Robustness in Machine Learning", Synthese, 2023, 202, Article 109. https://doi.org/10.1007/s11229-023-04334-9
12. Sejnowski T., Gorman R., "Connectionist Bench (Sonar, Mines vs. Rocks)", UCI Machine Learning Repository, 1988. https://doi.org/10.24432/C5T01Q
13. Sigillito V., Wing S., Hutton L., Baker K., "Ionosphere", UCI Machine Learning Repository, 1989. https://doi.org/10.24432/C5W01B
14. Wolberg W., Mangasarian O., Street N., Street W., "Breast Cancer Wisconsin (Diagnostic)", UCI Machine Learning Repository, 1993. https://doi.org/10.24432/C5DW2B
15. Little M., "Parkinsons", UCI Machine Learning Repository, 2007. https://doi.org/10.24432/C59C74
16. Hopkins M., Reeber E., Forman G., Suermondt J., "Spambase", UCI Machine Learning Repository, 1999. https://doi.org/10.24432/C53G6X
17. Breiman L., "Random Forests", Machine Learning, 2001, 45 (1), 5–32. https://doi.org/10.1023/A:1010933404324