

Explainability-Guided Feature Selection Model (XGFS-ML)

Sayera Yousufa¹, Jakkula Jayanthi², Racha Siri Varsha³, Omkar Radhika⁴

^{1,2,3}Computer Science Engineering Student, JNTUH, Telangana, India

⁴Omkar Radhika, Guest lecturer, Dept. of Computer Science and Engineering, JNTUH, Telangana, India

Abstract - Machine learning models often achieve high predictive accuracy but lack interpretability, limiting their applicability in critical domains. This paper proposes an Explainability-Guided Feature Selection Model (XGFS-ML), which integrates SHAP-based feature importance with supervised learning algorithms. A dynamic threshold-based feature selection mechanism is introduced to adaptively select relevant features based on their contribution to model predictions. The framework is evaluated using two datasets: the Breast Cancer Wisconsin dataset and a Telecom Churn dataset, ensuring improved generalization across domains. Performance evaluation is conducted using 5-fold cross-validation, reporting mean accuracy and standard deviation for robustness. Experimental results demonstrate that the proposed approach reduces feature dimensionality while maintaining or improving classification performance, particularly enhancing results in high-dimensional datasets. The framework provides a balanced solution for building accurate, interpretable, and efficient machine learning systems.

Key Words: Machine Learning, Feature Selection, XGFS-ML, Random Forest, Logistic Regression, Explainable AI, Classification.

1. INTRODUCTION

Machine Learning (ML) has become an essential technology for solving complex real-world problems across multiple domains including healthcare, finance, agriculture, cybersecurity, and intelligent automation [1], [9]. With the rapid growth of data availability, machine learning models are capable of learning complex patterns and relationships between input variables and target outputs.

However, many high-performance machine learning models operate as black-box systems, making their decision-making process difficult to interpret [10]. This lack of transparency reduces user trust, especially in critical applications where understanding model behavior is essential.

Feature selection is a crucial step in machine learning that aims to identify the most relevant features from a dataset. By eliminating irrelevant or redundant features, feature selection improves model accuracy, reduces computational complexity, and enhances generalization performance [4]. However, traditional feature selection techniques such as filter, wrapper, and embedded methods primarily focus on improving performance while often ignoring interpretability.

To address this limitation, this paper proposes an Explainability-Guided Feature Selection Model (XGFS-ML) that integrates SHAP-based feature importance with supervised machine learning algorithms. A dynamic threshold-based mechanism is introduced to adaptively select important features based on their contribution to model predictions.

Furthermore, to ensure robustness and generalization, the proposed framework is evaluated on multiple datasets, including a breast cancer dataset [6] and a telecom churn dataset. Model performance is assessed using stratified 5-fold cross-validation, with results reported as mean accuracy and standard deviation.

The proposed approach enhances both interpretability and performance, providing an effective solution for building transparent and efficient machine learning models.

1.1 Problem Statement

Many machine learning models achieve high predictive accuracy but lack interpretability. In sensitive domains such as healthcare and finance, it is important to understand how input features influence model decisions. Traditional feature selection approaches do not incorporate explainability into the selection process, which limits their usefulness in transparent decision-making systems.

1.2 Objectives of the study

The main objectives of this research are:

1. To develop an explainability-guided feature selection framework.
2. To integrate SHAP-based feature importance with machine learning algorithms.
3. To reduce feature dimensionality while maintaining or improving prediction accuracy.
4. To improve transparency and interpretability of machine learning models.

2. PROPOSED METHODOLOGY

2.1 System Architecture

The proposed framework consists of the following stages:

Dataset Collection → Data Preprocessing → SHAP-Based Feature Selection → Model Training → Performance Evaluation → Model Explanation

Initially, datasets are collected and preprocessed to handle missing values, encode categorical features, and standardize the data. Feature selection is then performed using SHAP-based importance scores. A dynamic threshold mechanism is applied to select the most relevant features based on their contribution to model predictions.

The selected features are used to train multiple machine learning models, including Logistic Regression, Random Forest, and XGBoost. Model performance is evaluated using stratified 5-fold cross-validation, and results are reported using mean accuracy and standard deviation to ensure robustness.

SHAP explainability is further used to analyze feature contributions, providing insights into model behavior. The pipeline enhances interpretability, reduces feature dimensionality, and maintains strong performance, making it suitable for real-world use.

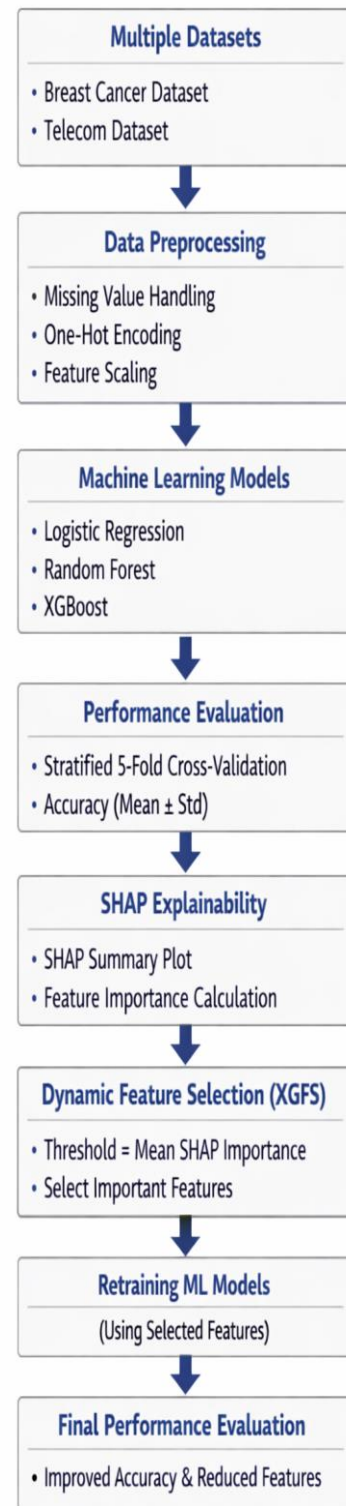


Fig 1 - Proposed XGFS-ML Architecture

2.2 Dataset Description

Two datasets were used in this study to evaluate the generalization capability of the proposed model.

1. Breast Cancer Wisconsin Dataset:

Source: UCI Machine Learning Repository [6]

Instances: 569

Features: 30 numerical features

Task: Binary classification (0 = Benign/1 = Malignant)

2. Telecom Churn Dataset:

Source: Public telecom customer dataset (Kaggle)

Features: Customer-related attributes such as tenure, monthly charges, contract type, payment method, and internet services

Task: Predict customer churn (Yes/No)

2.3 Data Preprocessing

Data preprocessing is an essential step in preparing the dataset for machine learning algorithms. The following preprocessing techniques were applied:

Handling missing values by replacing them with appropriate statistical measures

Encoding categorical variables using one-hot encoding
Feature scaling using standardization

Unlike traditional approaches, the dataset was not split using a fixed train-test ratio. Instead, stratified 5-fold cross-validation was used during model training and evaluation to ensure robust and unbiased performance estimation.

These preprocessing steps ensure that the machine learning models can effectively learn patterns while maintaining consistency and reliability in results [9].

2.4 Feature Selection

Feature selection is performed using SHAP (Shapley Additive Explanations) [5]. A dynamic threshold is computed based on the mean SHAP importance values, and features exceeding this threshold are selected. This adaptive approach ensures that feature selection is data-driven and flexible across different datasets. A minimum feature constraint is also applied to prevent excessive dimensionality reduction.

2.5 Machine Learning Algorithms

The following machine learning algorithms are used for performance evaluation:

Logistic Regression-

Logistic Regression is a statistical model used for binary classification tasks. It estimates the probability that an instance belongs to a specific class [1].

Random Forest-

Random Forest is an ensemble learning algorithm that constructs multiple decision trees and combines their predictions to improve accuracy and reduce overfitting [2].

XGBoost-

XGBoost is an advanced gradient boosting algorithm designed for efficiency and high performance in large-scale machine learning tasks [3].

3. EXPERIMENTAL RESULTS

The proposed XGFS-ML framework was evaluated using multiple machine learning models, including Logistic Regression, Random Forest, and XGBoost, on two datasets: Breast Cancer and Telecom Churn. The experiments were conducted to analyze the effectiveness of explainability-guided feature selection in improving classification performance while reducing feature dimensionality.

The models were initially trained using the full feature set. SHAP-based feature importance was then used to rank features, and a dynamic threshold mechanism was applied to select the most relevant features. The models were retrained using the optimized feature subset. Performance evaluation was carried out using stratified 5-fold cross-validation, and results are reported as mean accuracy along with standard deviation to ensure robustness and reliability. The results demonstrate that the proposed approach improves model performance while enhancing interpretability and reducing complexity.

3.1 Performance Metrics

The following evaluation metric was used to measure model performance:

- Accuracy – proportion of correctly classified samples.

To ensure reliable evaluation, stratified 5-fold cross-validation was used, and results are reported as mean accuracy \pm standard deviation. This provides a more robust assessment compared to a single train-test split.

3.2 Model Performance Comparison

The classification performance of different machine learning models and datasets is summarized in Table 1 and 2.

Table -1: Breast Cancer Dataset

MODEL	BEFORE XGFS	AFTER XGFS
LOGISTIC REGRESSION	0.9737 ± 0.0166	0.9789 ± 0.0131
RANDOM FOREST	0.9561 ± 0.0123	0.9596 ± 0.0153
XGBOOST	0.9684 ± 0.0163	0.9631 ± 0.0151

Table -2: Telecom Dataset

MODEL	BEFORE XGFS	AFTER XGFS
LOGISTIC REGRESSION	0.7745 ± 0.0162	0.8018 ± 0.0117
RANDOM FOREST	0.7346 ± 0.0002	0.7995 ± 0.0070
XGBOOST	0.8002 ± 0.0110	0.8002 ± 0.0113

The results show that the proposed XGFS framework maintains stable performance on the breast cancer dataset while improving performance in the telecom dataset. Overall, the approach enhances efficiency and interpretability across different datasets, demonstrating good generalization.

3.3 Model Accuracy Comparison

The accuracy comparison graph illustrates the performance of the machine learning models on the breast cancer dataset before and after applying the proposed XGFS feature selection approach.

The results indicate that Logistic Regression achieved an accuracy of 0.9737 before XGFS and improved to 0.9789 after feature selection, showing a noticeable performance gain. Random Forest achieved 0.9561 before XGFS and 0.9596 after XGFS, indicating a slight improvement. XGBoost achieved 0.9684 before XGFS and 0.9631 after XGFS, showing a minor decrease while still maintaining competitive performance.

Overall, the results demonstrate that the proposed XGFS approach effectively reduces feature dimensionality while maintaining or slightly improving model performance. Logistic Regression showed the most improvement, indicating that feature selection helped enhance model efficiency without compromising accuracy.

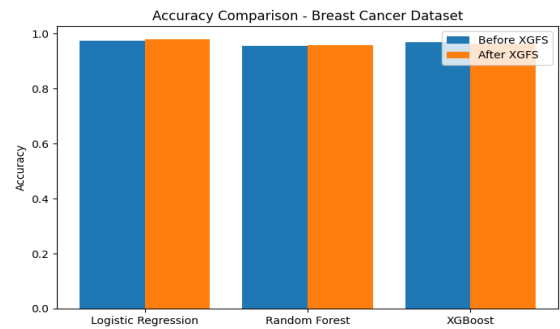


Fig -1: Model Accuracy Comparison

3.4 SHAP Feature Importance Analysis

To understand how individual features influence prediction outcomes, SHAP (Shapley Additive Explanations) [5] was used to compute feature importance scores.

The SHAP summary plot illustrates the impact of each feature on model predictions, where features are ranked based on their importance and the spread indicates their influence. Features such as area_worst, concave_points_worst, concave_points_mean, perimeter_worst, and concavity_worst are observed to have the highest impact on classification.

Higher feature values (shown in red) generally contribute positively toward the prediction, while lower values (shown in blue) have a negative impact. This indicates how feature variations affect the model output.

A dynamic threshold based on mean SHAP importance was applied to select the most relevant features, ensuring a data-driven and adaptive feature selection process. This improves interpretability while maintaining model performance.

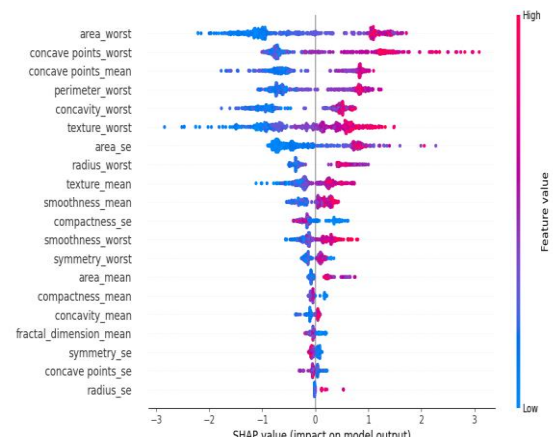


Fig -2: SHAP Summary Plot

3.5 Feature Importance Ranking

The feature importance graph ranks the most influential features based on their SHAP importance values. It highlights the relative contribution of each feature to the prediction output.

The analysis shows that features such as area_worst, concave_points_worst, concave_points_mean, perimeter_worst, and concavity_worst have the highest importance, indicating their strong influence on classification. These features consistently appear at the top of the ranking and play a key role in model decision-making.

By applying SHAP-based feature ranking along with a dynamic threshold, the proposed XGFS-ML framework effectively removes less significant features while retaining the most informative ones. This leads to improved interpretability and reduced dimensionality while maintaining strong classification performance.

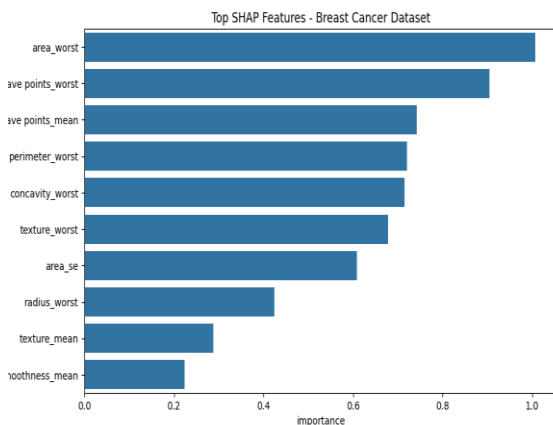


Fig -3: Top SHAP Feature Importance

4. CONCLUSIONS

This paper proposed an Explainability-Guided Feature Selection framework (XGFS-ML) that integrates SHAP-based feature importance with machine learning models using a dynamic threshold-based selection mechanism. The approach identifies the most influential features and removes redundant attributes, reducing dataset dimensionality while maintaining model performance [4].

Experimental evaluation on multiple datasets using Logistic Regression, Random Forest, and XGBoost demonstrates that the proposed method maintains or improves predictive accuracy while enhancing model efficiency. Notable improvements were observed in the telecom dataset, highlighting the effectiveness of the approach in handling high-dimensional real-world data.

SHAP analysis further provided insights into feature contributions, improving transparency and interpretability of the models. Overall, the proposed framework offers a

reliable and efficient solution for building interpretable machine learning systems.

Future work will focus on applying the framework to large-scale datasets and exploring advanced learning techniques.

REFERENCES

- [1] T. Mitchell, "Machine Learning," McGraw-Hill, 1997.
- [2] L. Breiman, "Random Forests," Machine Learning Journal, vol. 45, no. 1, pp. 5–32, 2001.
- [3] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.
- [4] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," Journal of Machine Learning Research, vol. 3, pp. 1157–1182, 2003.
- [5] S. Lundberg and S. Lee, "A Unified Approach to Interpreting Model Predictions," Advances in Neural Information Processing Systems (NeurIPS), 2017.
- [6] D. Dua and C. Graff, "UCI Machine Learning Repository," University of California, Irvine, 2017.
- [7] J. Han, M. Kamber, and J. Pei, "Data Mining: Concepts and Techniques," Morgan Kaufmann Publishers, 2011.
- [8] A. Géron, "Hands-On Machine Learning with Scikit-Learn, Keras and TensorFlow," O'Reilly Media, 2019.
- [9] P. Domingos, "A Few Useful Things to Know About Machine Learning," Communications of the ACM, vol. 55, no. 10, pp. 78–87, 2012.