

Developing Kernel-Aware Explainability Techniques for Interpretable Support Vector Machines

OSUNNIYI SEGUN JAMES¹, YUSUF MARIAM AYOMIDE²

^{1,2}*Department of Computer Science, Ladoke Akintola University of Technology, Oyo State. Nigeria*

Abstract- Artificial Intelligence (AI) systems are increasingly used for classification and decision support, but many machine learning models provide limited insight into how predictions are generated. This problem is particularly important in sensitive domains where users need to understand and trust automated decisions. Support Vector Machines (SVMs) provide strong classification performance, but nonlinear kernel functions make their decision processes difficult to interpret. This study developed four kernel-aware explainability techniques for providing insights into SVM decision-making: Support Vector Proximity Visualization (SVPV), Kernel Contribution Decomposition (KCD), Feature Influence Plot (FIP), and Decision Region Mapping (DRM). The Breast Cancer Wisconsin Diagnostic Dataset was used to evaluate three SVM models based on Linear, Polynomial, and Radial Basis Function (RBF) kernels. The models achieved testing accuracies of 95.2%, 96.0%, and 97.4%, respectively, while ROC-AUC scores were 0.96, 0.97, and 0.99. SVPV provided information about the relationship between test observations and influential support vectors, KCD examined kernel and support-vector contributions, FIP identified influential features, and DRM visualized classification regions and decision boundaries. Comparison with LIME and SHAP indicated that the proposed techniques provided high interpretability and stability with moderate computational requirements. User evaluation produced mean scores of 4.3, 4.1, 4.5, and 4.4 out of 5 for SVPV, KCD, FIP, and DRM, respectively. The findings demonstrate that model-specific, kernel-aware explanations can improve the transparency of SVM-based classification while retaining strong predictive performance.

Keywords: Explainable Artificial Intelligence, Support Vector Machines, Kernel Methods, Interpretability, Feature Influence, Decision Visualization.

I. INTRODUCTION

Artificial Intelligence (AI) and machine learning have become important tools for classification, prediction, and decision support. Despite their effectiveness, many machine learning models function as black

boxes, producing predictions without providing sufficient information about how those predictions were obtained. This lack of transparency can reduce user trust and make model validation and error analysis difficult [1].

Support Vector Machines (SVMs) are established supervised learning algorithms that construct decision boundaries by maximizing the margin between classes [2]. Their ability to use kernel functions makes them effective for nonlinear classification. Linear, Polynomial, and Radial Basis Function (RBF) kernels allow SVMs to model relationships of increasing complexity. However, this flexibility introduces an interpretability problem because nonlinear transformations are performed implicitly.

Previous studies have examined methods for interpreting SVMs, including support-vector analysis and rule extraction [3]. More general Explainable Artificial Intelligence (XAI) techniques, such as LIME and SHAP, provide local and feature-level explanations for black-box models [4], [5]. However, model-agnostic methods do not necessarily expose the specific roles of support vectors and kernel functions in SVM decisions.

This study addresses this gap by developing four kernel-aware explanation techniques: Support Vector Proximity Visualization (SVPV), Kernel Contribution Decomposition (KCD), Feature Influence Plot (FIP), and Decision Region Mapping (DRM). The techniques provide complementary views of SVM decisions at support-vector, kernel, feature, and decision-region levels. The Breast Cancer Wisconsin Diagnostic Dataset was used to evaluate the techniques.

II. II LITERATURE REVIEW

This section reviews existing studies related to explainable artificial intelligence, support vector machines, existing explainability techniques, and related support vector machine approaches.

A. Explainable Artificial Intelligence

Explainable Artificial Intelligence concerns methods that enable users to understand the behaviour and outputs of AI systems. Interpretability is important when machine learning decisions affect individuals or require human validation. Doshi-Velez and Kim emphasized the need for rigorous approaches to evaluating interpretability [1]. XAI methods can be intrinsic, where interpretability is incorporated into the model, or post-hoc, where explanations are generated after model training [6].

B. Support Vector Machines

SVMs classify observations by identifying an optimal decision boundary with maximum separation between classes. Observations closest to the boundary are support vectors and play a central role in defining the classifier [2]. Kernel functions extend SVMs to nonlinear problems. While linear SVMs provide relatively direct feature relationships, nonlinear kernels make interpretation more difficult because the transformed feature space is not directly observable.

C. Existing Explainability Techniques

LIME explains individual predictions by constructing a simpler local approximation around a selected observation [4]. SHAP provides feature-attribution values based on Shapley values and offers a unified framework for interpreting model predictions [5]. These methods are broadly applicable but are not specifically designed to expose SVM support-vector or kernel behaviour.

D. Research Gap

Existing SVM interpretation approaches generally focus on support vectors, feature weights, or extracted rules, while general XAI methods focus on model-agnostic local or feature-level explanations. Consequently, there remains a need for explanation techniques that explicitly combine support-vector relationships, kernel contributions, feature influence,

and decision-region information. The proposed framework addresses this gap through four complementary visualization and analysis techniques.

E. Related Works

Existing research demonstrates substantial interest in explainability, but comparatively fewer approaches focus specifically on the internal structure of kernel-based SVMs. Support-vector analysis identifies influential observations but does not directly explain feature-level contributions. Linear feature-weight analysis is useful for linear SVMs but becomes less straightforward for nonlinear kernels. Dimensionality-reduction methods provide visual representations but may not explain the actual computational contribution of individual components.

The reviewed literature therefore indicates four major gaps: limited kernel-specific local explanations, difficulty visualizing high-dimensional decision behaviour, limited interpretation of nonlinear kernel contributions, and insufficiently user-oriented explanation tools. The present study addresses these gaps by combining support-vector analysis, kernel contribution analysis, feature perturbation, and decision-region visualization into a single explainability framework.

III. METHODOLOGY

A. Research Design and Dataset

The study adopted a Design Science Research approach involving problem identification, design, implementation, demonstration, evaluation, and refinement of explainability techniques. The Breast Cancer Wisconsin Diagnostic Dataset was selected for experimentation. It contains 569 observations and 30 numerical diagnostic features describing characteristics of cell nuclei obtained from digitized breast-mass images. The observations belong to malignant and benign classes.

The dataset was divided into training and testing subsets using a 70:30 partition. Features were normalized between 0 and 1 before model development.

B. SVM Model Development

Three SVM classifiers were implemented using Linear, Polynomial, and RBF kernels. The Polynomial kernel used degree 3. The models were implemented in Python using Scikit-learn, with NumPy and Pandas supporting data processing and Matplotlib and Seaborn supporting visualization.

The RBF kernel was used to represent nonlinear relationships through similarity between observations:

$$K(x_i, x_j) = - \|x_i - x_j\|^2$$

The three models provided a basis for examining how kernel choice affects both classification performance and interpretability.

C. Proposed Explainability Techniques

Support Vector Proximity Visualization (SVPV): SVPV identifies and visualizes the relationship between a selected test observation and influential support vectors. It provides local information about observations associated with a particular prediction.

Kernel Contribution Decomposition (KCD): KCD analyzes the contributions of kernel operations and influential support vectors to an SVM prediction. It was designed to expose information that is normally hidden within nonlinear kernel computations.

Feature Influence Plot (FIP): FIP evaluates feature influence by perturbing individual features and examining changes in the model's decision function. Features can consequently be ranked and visualized according to their influence on a prediction.

Decision Region Mapping (DRM): DRM visualizes the structure of the SVM decision space. Principal Component Analysis (PCA) was used to project the high-dimensional dataset into two dimensions, allowing decision regions, test observations, support vectors, and margins to be displayed.

The four techniques provide complementary explanations as shown in Table 1.

Table 1: Proposed Explainability Techniques

Technique	Explanation Focus
SVPV	Support-vector proximity
KCD	Kernel/support-vector contribution
FIP	Feature influence
DRM	Decision regions and boundaries

D. Evaluation

Predictive performance was evaluated using classification accuracy, confusion matrices, and ROC-AUC. Explanation techniques were evaluated according to computational requirements, stability, fidelity, clarity, visual quality, and usefulness. LIME and SHAP were used as comparison methods. User evaluation was also conducted using a five-point scale.

IV. RESULTS

This section presents the implementation results, system performance evaluation, security analysis, and discussion of findings obtained from the developed cloud-based Personnel Management Information System (PMIS).

A. SVM Classification Performance

The three SVM models produced high classification performance.

Table 2: SVM Classification Performance

Kernel	Training Accuracy (%)	Testing Accuracy (%)	ROC-AUC
Linear	96.4	95.2	0.96
Polynomial	97.1	96.0	0.97
RBF	98.3	97.4	0.99

The RBF model achieved the highest training accuracy of 98.3%, testing accuracy of 97.4%, and ROC-AUC of 0.99. The Polynomial model achieved 96.0% testing accuracy and 0.97 ROC-AUC, while the Linear model achieved 95.2% and 0.96, respectively.

The results indicate that nonlinear kernels captured relationships in the dataset more effectively than the linear kernel. The relatively small difference between training and testing accuracy for the RBF model also indicates strong generalization within the experimental dataset.

B. Explainability Results

SVPV provided a visual representation of test instances and nearby support vectors, making it possible to identify observations associated with local classification behaviour. Instances located closer to the decision boundary could consequently be examined more closely.

KCD exposed the contributions of influential support vectors and kernel computations. The results showed that only a subset of support vectors contributed substantially to particular predictions, providing a more focused representation of the SVM decision process.

FIP identified influential diagnostic features through perturbation analysis. Features including mean radius, mean texture, mean perimeter, mean area, concavity, and concave points were among the influential variables identified during the analysis. FIP provided a relatively simple feature-level explanation of individual predictions.

DRM used PCA to project the original high-dimensional data into two dimensions and displayed the resulting decision regions. The visualization showed that the RBF model generated more complex nonlinear regions than the Linear model, reflecting the greater flexibility of the RBF kernel.

C. Comparison with LIME and SHAP

Table 3: Comparison of Explanation Methods

Criterion	LIME	SHAP	Proposed Methods
Local explanation	Yes	Yes	Yes
Interpretability	Medium	High	High
Stability	Low	Medium	High
Computational requirement	High	High	Moderate
Kernel-specific information	Limited	Limited	High

LIME and SHAP are powerful general-purpose XAI methods [4], [5], but their model-agnostic nature limits their ability to directly expose SVM-specific structures. The proposed methods provide information

specifically related to support vectors, kernel contributions, feature influence, and decision regions. The combination of the techniques therefore provides broader SVM-specific insight than a single generic explanation method.

D. User Evaluation

Table 4: Mean User Evaluation Scores

Technique	Mean Score / 5
SVPV	4.3
KCD	4.1
FIP	4.5
DRM	4.4

All four techniques received positive evaluations. FIP achieved the highest mean score of 4.5, followed by DRM at 4.4 and SVPV at 4.3. KCD recorded 4.1.

The relatively high FIP score suggests that feature-level explanations were particularly easy for users to understand. DRM also received a high score because visual representations of decision regions provide an intuitive overview of classification behaviour. KCD's slightly lower score may be associated with the greater conceptual complexity of explaining kernel contributions.

Overall, the results indicate that the proposed techniques can provide useful explanations without requiring users to understand the complete mathematical structure of an SVM.

V. CONCLUSION

This study developed four kernel-aware explainability techniques for interpreting Support Vector Machine decision-making: Support Vector Proximity Visualization, Kernel Contribution Decomposition, Feature Influence Plot, and Decision Region Mapping. Three SVM models based on Linear, Polynomial, and RBF kernels were evaluated using the Breast Cancer Wisconsin Diagnostic Dataset. The RBF model achieved the strongest classification performance, with 97.4% testing accuracy and 0.99 ROC-AUC. The proposed explanation techniques complemented this predictive performance by providing information

about support-vector relationships, kernel contributions, feature influence, and decision regions. Comparison with LIME and SHAP showed that the proposed methods offered high interpretability and stability with moderate computational requirements. User evaluation also produced positive results, with scores ranging from 4.1 to 4.5 out of 5.

The study demonstrates that model-specific explainability techniques can improve the transparency of kernel-based SVMs while preserving their predictive capability. Future work should evaluate the techniques on larger and more diverse datasets, investigate their application to other machine learning architectures, develop interactive real-time explanation systems, and conduct validation with appropriate domain professionals.

REFERENCES

- [1] F. Doshi-Velez and B. Kim, “Towards a rigorous science of interpretable machine learning,” arXiv preprint arXiv:1702.08608, 2017.
- [2] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.
- [3] N. H. Barakat and A. P. Bradley, “Rule extraction from support vector machines: A review,” *Neurocomputing*, vol. 74, no. 1–3, pp. 178–190, 2010.
- [4] M. T. Ribeiro, S. Singh, and C. Guestrin, ““Why should I trust you?”: Explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016.
- [5] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems*, vol. 30, pp. 4765–4774, 2017.
- [6] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, “A survey of methods for explaining black box models,” *ACM Computing Surveys*, vol. 51, no. 5, pp. 1–42, 2018.