Scientific Journal of Informatics
Vol. 13, No. 2, May 2026
p-ISSN 2407-7658 https://journal.unnes.ac.id/journals/sji/index e-ISSN 2460-0040

A Comparative Evaluation of Model-Based and SHAP-Based Feature Analysis for Robust Health Data Classification

Indra Waspada[1*], Satriawan Rasyid Purnama[2], Alwey Hakim[3], Alfonso Clement Sutantio[4]
[1, 2, 3,4]Department of Informatics, Universitas Diponegoro, Indonesia

Abstract.
Purpose: Feature selection is one critical element of healthcare data classification, which directly affects predictive performance, model robustness, and interpretability. Nevertheless, traditional model-based feature importance methods are unstable in robustness and provide random or misleading results on high-dimensional and heterogeneous healthcare data. The purpose of this paper was to assess model-based and SHapley Additive exPlanations (SHAP) - based feature analysis on multimodal healthcare data classification in a comparative manner.
Methods: This study employed a quantitative comparative experimental design using real Electronic Medical Records (EMR) data captured from primary care clinics. The dataset comprises 2,158 patient records, with numerical and textual features in a multimodal feature space. The analytical pipeline included data acquisition, preprocessing, multimodal feature integration, and model development using six supervised learning algorithms from ensemble-based and margin-based categories. The model-based feature importance was used with the SHAP-based feature importance. We examined the robustness of this method across different scenarios through systematic feature ablation using baseline, strong, and weak features.
Result: Experimental results demonstrate that model-based feature importance exhibits unpredictable behavior when features are removed and is sometimes counterintuitive. On the other hand, feature importance based on SHAP is consistent and linearly proportional across the models under consideration. The highest macro F1-score was achieved by Extra Trees with SHAP-based strong features (0.824), exceeding the baseline (0.811). In robustness testing, SHAP-based weak-feature removal reduced the LightGBM F1-score from 0.764 to 0.575, suggesting a clearer distinction between informative and non-informative features. Overall, SHAP-based feature selection provided a more reliable and interpretable framework for multimodal healthcare classification.
Novelty: Our experiment results demonstrate empirically that the SHAP-based feature importance is more robust and reliable than conventional model-based approaches when it comes to extracting features from multimodal medical records. This work shows that SHAP is more than a post hoc explanation by presenting it as an interpretable feature selection criterion guiding feature relevance analysis in healthcare machine learning.

Keywords: Feature Importance Analysis, SHAP, Multimodal, Health Data Classification, Robustness Analysis
Received February 2026 / **Revised** April 2026 / **Accepted** April 2026

INTRODUCTION

The introduction of Electronic Medical Records (EMR) has resulted in a growing demand for the health, condition classification systems to underpin clinical decision making [1], [2], [3], [4]. Supervised machine learning models have been widely used to deal with medical records [5], [6], [7], [8]. Notwithstanding, predictive accuracy of model alone is not enough; interpretability, transparency and usability are also essential [4], [9], [10].

Patient-centered data is inherently heterogeneous, comprising structured numeric measurements like heart rate, and blood pressure with textually recorded descriptions of symptoms or complaints [11]. These multimodal approaches have shown that performance can be further improved by using different types of data together. Nevertheless, this concatenation produces a high dimensional and unbalanced feature space [4], [8], [12], [13]. However, this make it difficult to interpret the contribution of each feature on their role in classification especially in patient safety-critical clinical scenario [8].

*Corresponding author.
Email addresses: indrawaspada@lecturer.undip.ac.id (Waspada)*
DOI: 10.15294/sji.v13i2.42406

To answer this challenge, feature importance analysis is generally used. Many existing studies rely on model-based feature importance derived from internal learning mechanisms, such as impurity-based measures or coefficient magnitudes [14], [15], [16]. While computationally efficient, these approaches provide only global rankings and are highly dependent on the learning algorithm. In heterogeneous clinical data, such dependency may lead to unstable or counter-intuitive interpretations, where features identified as important do not consistently correspond to features that genuinely influence predictive outcomes [17], [18].

Explainable machine learning has emerged as an important paradigm for providing prediction-centric explanations, with SHapley Additive exPlanations (SHAP) offering a theoretically grounded framework for attributing feature contributions at both global and local levels [9], [19], [20], [21], [22]. Although SHAP has become one of the most widely adopted techniques for interpreting predictive models, particularly in healthcare applications, its use is predominantly limited to post hoc explanation and visualization [23]. However, limited research has investigated SHAP as a systematic basis for feature relevance analysis and feature selection, particularly in comparison with conventional model-based importance measures [24], [25], [26]. Furthermore, robustness evaluations of model-based and SHAP-based features have also been limited, despite the critical importance of robustness, particularly in the healthcare domain [23], [27], [28].

Based on this gap, this study proposes a controlled comparative analysis of model-based (conventional) feature analysis and SHAP-based feature analysis. The main contributions of this research are as follows:

1. Conducting a comprehensive controlled comparative analysis of importance features based on supervised learning models that include ensemble-based and margin-based models on multimodal clinical medical record data.
2. Performing feature analysis using the same classification models using SHAP-based importance features.
3. Comparing between model-based and SHAP-based approaches by applying robustness evaluation using systematic feature ablation to assess consistency and reliability of feature relevance.

METHODS

This study adopts the CRISP-ML framework [10], [29] to support controlled and reproducible comparisons as illustrated in Figure 1. This approach exploits two parallel experimental paths that are exactly the same going from how we treat the dataset through pre-processing, learning algorithms and evaluation settings. This makes it possible to separate the contributions of feature analysis strategies toward prediction accuracy, stability and interpretability.

Business and Data Understanding

We present an application of PPS in the primary care clinic setting by analysing clinical records collected at a primary healthcare clinic (Clinic X) during August–September 2023 based on routine patient management data [30]. This dataset contains 2,158 patients with 16 clinical measurement attributes, diagnostic codes and symptom narratives, in raw Indonesian text. Such features can be either numerical clinical variables (e.g., blood pressure, weight, height and age) or categorical or textual ones (e.g., gender, address, diagnosis, therapy and symptoms), resulting in multimodal representation of data for analysis. A list of these features and their properties is given in Table 1.

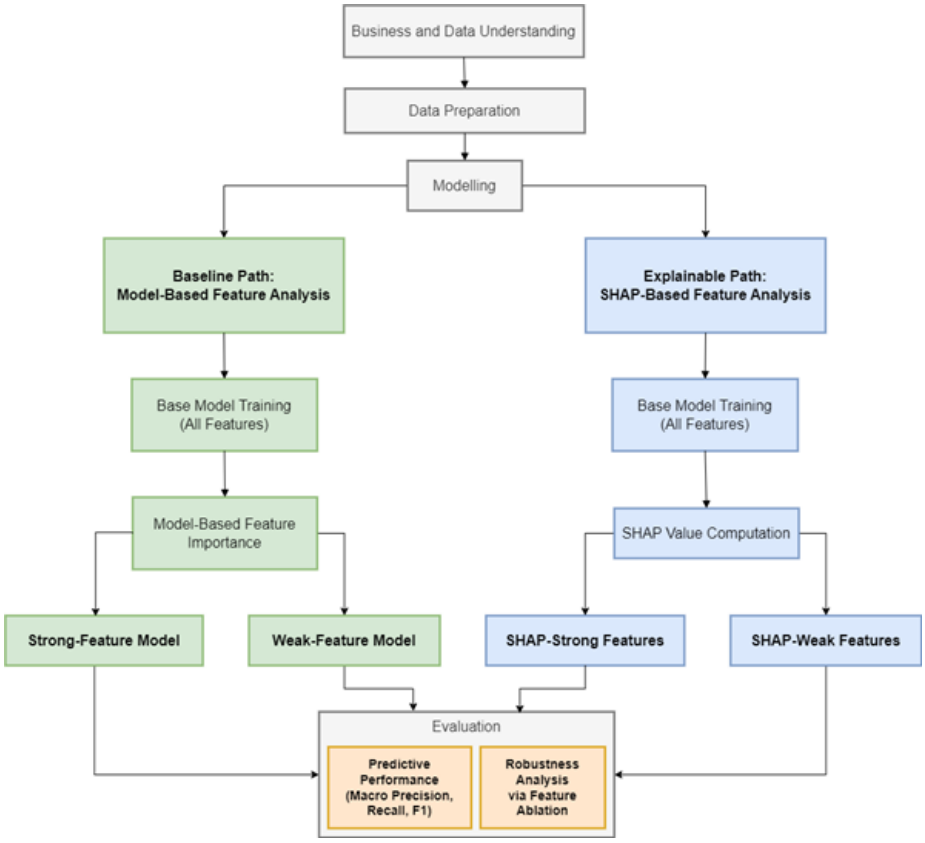

Figure 1. CRISP-ML-Based Workflow for Comparative Evaluation of Model-Based and SHAP-Based Feature Analysis

Table 1. Summary of Clinical Dataset Attributes

No.	Attribute	Data Type	Description
1	NO	Numeric	Record index
2	NO_RM	String	Unique medical record identifier
3	NAME	String	Patient name
4	GENDER	String	Patient gender
5	ADDRESS	String	Sub-district or district
6	RT	Numeric	Neighborhood association code
7	RW	Numeric	Community association code
8	DIAGNOSES	String	**Diagnosis based on ICD-10**
9	THERAPI	String	Prescribed medication
10	SYMPTOMS	String	**Free-text symptom description**
11	SYSTOLE	Numeric	Systolic blood pressure
12	DIASTOLE	Numeric	Diastolic blood pressure
13	PULSE	Numeric	Pulse rate
14	WEIGHT	Numeric	Body weight
15	HEIGHT	Numeric	Height
16	AGE	Numeric	Patient age

Owing to data limitations and class imbalance, we consider five diagnostic classes based on the most prevalent ICD-10 categories; less commonly occurring diagnoses are grouped as an Other class [30]. This cluster is designed to capture the typical outpatient disease while ensuring enough sample size for each class. The resulting diagnostic groups are presented in Table 2.

Table 2. Diagnostic Classes Derived from ICD-10 Codes

Class	ICD-10 Category Description
Class J	Diseases of the respiratory system
Class R	Symptoms, signs, and abnormal clinical findings not elsewhere classified
Class I	Diseases of the circulatory system
Class K	Diseases of the digestive system
Others	Remaining ICD-10 disease categories

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ be the entire dataset considered in this work, where x_i is the multimodal feature vector of an i-th patient record and y_i is the associated diagnostic class label. This data collection has a multimodal topology of a few numerical dimensions and high-dimensional text like features. The text attributes can be a conversion to numerical representation using which, the feature factor becomes a sparse vector space distributed inconspicuously. This feature is based on the natural composition of data of patient records in practical clinical environment.

Data Preparation

Data preparation is completed to convert raw patient medical record data into a structured representation that is amenable for supervised learning and comparative feature analysis. Preprocessing is performed independently on the numeric clinical features and textual symptom descriptions. This is then integrated into a single multimodal representation as depicted in Figure .

Continuous clinical characteristics were retained as numerical variables. Null or incorrect values were either replaced with the median using imputation [4], [5]. Z-score normalization was applied to have standardized scales between features of variables [8], [12]. Reported Diagnostic descriptions entered as text were transformed to numerical representations with term frequency–inverse document frequency (TF-IDF).

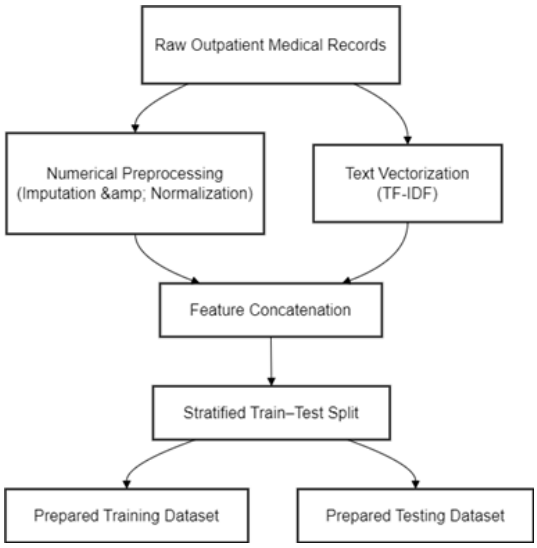

Figure 2. Data Preparation

It fuses the numeric features and textual features to form a unified multimodal feature representation [4], [31]. Any dimensionality reduction is not employed to keep the granularity of features for downstream feature importance analysis, i.e., SHAP-based attribution and feature ablation tests. Finally, the dataset is divided into training and testing data based on their class distribution by stratified sampling.

Modeling

The modeling stage is designed to enable a controlled comparison between model-based feature analysis and SHAP-based explainable feature analysis under identical experimental conditions. Six supervised learning algorithms are evaluated to represent diverse modeling paradigms commonly used in health data classification, including ensemble-based classifiers (Random Forest [32], Extra Trees [33], XGBoost [34], LightGBM [35])) and margin-based classifiers (Support Vector Machine with radial basis function kernel and Linear Support Vector Machine [36], [37]).

For each learning algorithm, two parallel feature analysis strategies are applied. In the model-based feature analysis approach, feature relevance is derived from model-internal importance measures, such as impurity-based scores for tree-based models and coefficient magnitudes for linear classifiers. In the SHAP-based feature analysis approach, feature relevance is estimated using SHAP values that quantify prediction-centric feature contributions. In both cases, features are ranked according to their relevance scores to construct strong-feature and weak-feature subsets.

Table 3. Hyperparameter Configuration of the Evaluated Machine Learning Models

Model	Hyperparameters	Value
Random Forest	Number of estimators	200
	Splitting criterion	Gini
	Class weighting	Balanced
	Number of parallel jobs	-1 (all cores)
Extra Trees	Number of estimators	200
	Splitting criterion	Gini
	Number of parallel jobs	-1 (all cores)
XGBoost	Number of estimators	200
	Maximum tree depth	Default
LightGBM	Number of estimators	200
	Class weighting	Balanced
SVM (RBF kernel)	Kernel type	RBF
	Regularization parameter (C)	1.0
	Class weighting	Balanced
Linear SVM	Regularization parameter (C)	1.0
	Penalty	L2
	Class weighting	Balanced

Models are retrained using the complete feature set as well as the derived strong-feature and weak-feature subsets to assess sensitivity to feature selection and robustness under feature removal. To ensure fairness and isolate the effect of feature analysis methodology, all models are trained using fixed hyperparameter configurations, which are summarized in Table 3. Hyperparameter tuning is intentionally omitted to avoid confounding effects and to maintain comparability across experimental settings.

Evaluation

The predictive performance of the models was assessed by considering precision, recall and macro-average F1-score (to mitigate class imbalance in dataset) [38], [39]. The evaluation was performed with three different feature settings, baseline, strong and weak features. The aim of these variations was to study the sensitivity in terms of performance with respect to model-based and SHAP-based feature selections.

Robustness is measured by feature removal, i.e., the performance drop after removing features which are deemed relevant by each method. Feature attribution methods where the drop in performance is consistent and proportional to removal of features are considered more trustworthy. The fact that non-constant responses are irregular suggests instability in estimating the relevance of features [40].

RESULTS AND DISCUSSIONS

This section compares feature analysis from model-based and SHAP-based methods. Emphasis is on prediction performance behavior and feature-removal robustness. The results iteratively analyzes performance consistency across models and feature sets to check if the discovered features relevance of individual method matches with the underlying dependencies of a model in classifying multimodal health data.

Table 4 shows the results of precision, recall and average macro F1-score for the baseline model. It seems that the ensemble method is probably better in macro-average performance than margin-based approach. This demonstrates the advantage of ensemble models in modeling heterogeneous and high-dimensional feature spaces. It can be seen that the ensemble approach provides more balanced performance on diagnostic classes. These findings are of particular significance in the context of imbalanced healthcare data, where minority (classes which tend to be underrepresented) are required to be correctly identified represented in order not to lose critical information.

Table 4. Macro-Averaged Recal, Precision, and F1-Score

Model	Precision	Recall	F1-Score
Random Forest	**0.848453**	0.778098	0.8058
Extra Trees	0.845298	0.786167	**0.8111**
XGBoost	0.784319	0.763623	0.7710
LightGBM	0.799168	0.740608	0.7638
RBF-SVM	0.716248	0.779201	0.7370
Linear SVM	0.819163	**0.810105**	0.8098

Robustness is checked through systematic feature ablation by analyzing performance changes after feature removal. Table 5 and Table 6 compare model-based and SHAP-based feature importance for feature selection across classifiers using precision and recall. Figure and Figure s show the macro-average F1 scores. These results show that the strong feature subset consistently outperforms both the baseline and weak features. However, a key finding is that SHAP-based selection yields more stable and balanced performance improvements, particularly in macro-average recall because SHAP ranks features according to their marginal contribution to prediction outcomes. Consequently, removing highly ranked features reduces discriminative information, whereas retaining strong features preserves predictive structure. This leads to a monotonic and interpretable performance response across classifiers.

In tree-based ensemble models, strong SHAP-based features generally produce higher recall with comparable precision, resulting in superior or equivalent macro F1-scores, with Extra Trees showing the best overall performance. These results indicate that SHAP better captures non-linear relationships and feature interactions than conventional impurity-based measures. Moreover, weak-feature subsets selected by SHAP cause sharper performance declines, suggesting clearer separation between informative and non-informative features. In contrast, model-based importance may become unstable because internal importance scores are algorithm-dependent and sensitive to multicollinearity, feature interactions, and sampling variation, particularly in multimodal data with correlated textual and numerical attributes.

For margin-based classifiers, particularly the RBF-SVM model based feature selection has a fluctuating performance. Yet SHAP-based selection might lead to a more stable value of macro F1 score. In general, we have shown that SHAP-based feature importance is a model-independent feature selection strategy. These findings suggest that SHAP-based feature selection is often more stable for improving performance than model-based methods in multi-class and multimodal classification tasks.

Table 5. Model-based Macro-Averaged Recall and Precision

Model	Precision			Recall		
	Base Feature	Strong Feature	Weak Feature	Base Feature	Strong Feature	Weak Feature
Random Forest	0.848453	**0.856599**	0.683194	0.778098	**0.783028**	0.631475
Extra Trees	0.845298	**0.860773**	0.66111	0.786167	**0.794512**	0.634808
XGBoost	0.784319	**0.812447**	0.788492	0.763623	**0.804666**	0.756916
LightGBM	0.799168	**0.804484**	0.70664	0.740608	**0.758040**	0.710226
RBF-SVM	0.716248	0.476706	**0.788276**	0.779201	0.533275	**0.808185**
Linear SVM	**0.819163**	0.806694	0.806629	0.810105	0.796854	**0.819255**

Table 6. SHAP-based Macro-Averaged Recall and Precision

Model	Precision			Recall		
	Base Feature	Strong Feature	Weak Feature	Base Feature	Strong Feature	Weak Feature
Random Forest	0.848453	**0.857439**	0.76942	0.778098	**0.793555**	0.685981
Extra Trees	0.845298	**0.859633**	0.802932	0.786167	**0.798146**	0.742416
XGBoost	0.784319	**0.784340**	0.722186	0.763623	**0.766761**	0.708180
LightGBM	0.799168	**0.813047**	0.660577	0.740608	**0.787367**	0.554847
RBF-SVM	0.716248	**0.730092**	0.633564	0.779201	**0.792233**	0.602725
Linear SVM	0.819163	**0.823995**	0.738230	0.810105	**0.805671**	0.763790

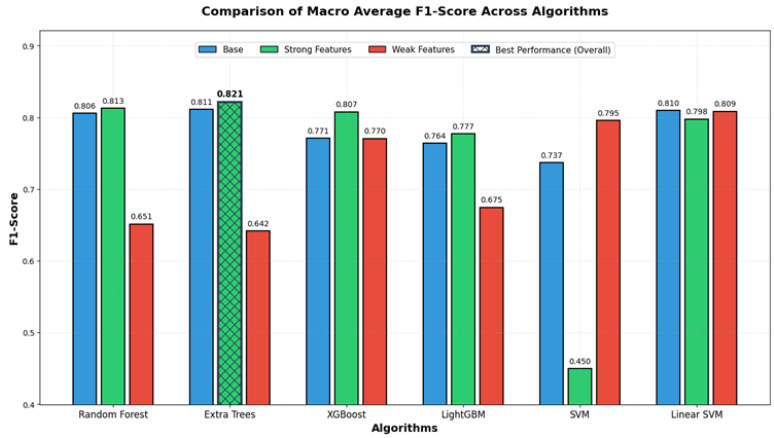

Figure 3. Model-based Macro-Average F1-score

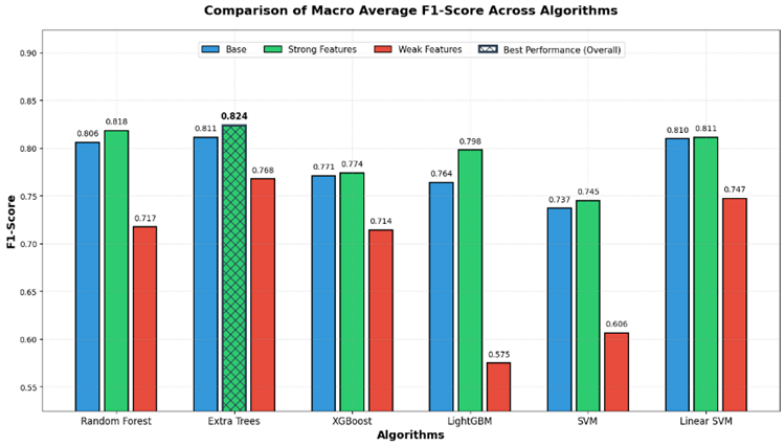

Figure 4. SHAP-based Macro-Average F1-score

In Table 7, we analyze classifier robustness through the consistent macro-F1 AUC variation comparison between baseline, strong and weak features subsets. Our findings demonstrate that feature significance by using SHAP is the reliable method to obtain the proportional robustness for all classifiers. These results show consistent and predictable performance behaviour as the quality of extracted features changes. Model-based feature selection on the other hand generates more varied, less sensitive or contradictive results over most models. The results of RBF-SVM are the most fluctuating, in which active weak features outperform passive strong features. Both tree-based ensemble and margin-based classifiers show clearer or more interpretable robustness patterns for SHAP-based selection. These results also provide further evidence that SHAP-based feature importance enables better robustness and reliability in selecting features.

Table 7. Robustness Response of Classifiers

Model	Approach	F1-Score			Robustness Response
		Base Feature	Strong Feature	Weak Feature	
Random Forest	Model-Based	0.8058	0.8127	0.6514	Inconsistent
	SHAP-Based	0.8058	0.8182	0.7173	**Proportional**
Extra Trees	Model-Based	0.8111	0.8214	0.6417	Inconsistent
	SHAP-Based	0.8111	0.8236	0.7678	**Proportional**
XGBoost	Model-Based	0.7710	0.8075	0.7704	Weak sensitivity
	SHAP-Based	0.7710	0.7740	0.7142	**Proportional**
LightGBM	Model-Based	0.7638	0.7772	0.6746	Inconsistent
	SHAP-Based	0.7638	0.7977	0.5747	**Proportional**
RBF-SVM	Model-Based	0.7370	0.4495	0.7955	Counter-intuitive
	SHAP-Based	0.7370	0.7449	0.6064	**Proportional**
Linear SVM	Model-Based	0.8098	0.7975	0.8085	Weak sensitivity
	SHAP-Based	0.8098	0.8111	0.7471	**Proportional**

The empirical findings discussed above are summarized in the comparative summary given in Table 8. It compares model-based and SHAP-based feature analysis in aspects of the methodology. Our comparison highlights that while model-based feature analysis serves as a reasonable baseline, SHAP-based feature analysis introduces more stable and interpretable behavior across models in the case of heterogeneous health data.

Table 8. Comparative Summary of Model-Based and SHAP-Based Feature Analysis

Aspect	Model-Based Feature Analysis	SHAP-Based Feature Analysis
Feature relevance basis	Model-internal importance	Prediction-based contribution
Behavior under weak feature removal	Inconsistent	Consistent performance degradation
Behavior under strong feature removal	Not always proportional	Clear and proportional degradation
Robustness pattern	Unstable across models	Stable across models
Ensemble model behavior	Moderately reliable	Highly reliable
Margin-based model behavior	Counter-intuitive (e.g., RBF-SVM)	More controlled and consistent
Suitability for health data	Baseline reference	High (trustworthy & explainable)

For multimodal healthcare datasets, reliable feature attribution is essential because clinical decisions often require transparent reasoning across heterogeneous variables such as symptoms, vital signs, and diagnoses. The superior stability of SHAP suggests greater suitability for trustworthy decision-support systems.

CONCLUSION

This study conducted a controlled comparative evaluation of conventional model-based and SHAP-based feature importance methods for multimodal healthcare data classification. Experimental results demonstrate that, while the two approaches have similar predictive capacity, the feature importance computed from models exhibits erratic, at times unintuitive behavior as features are omitted. On the other hand, SHAP-based feature analysis showed a gradual, consistent, and proportional decline in performance as informative features were excluded. These findings suggest that SHAP more accurately reflects feature relevance by closely capturing underlying model dependencies across classifiers. This study further highlights that interpretability and robustness are essential evaluation criteria alongside predictive accuracy in healthcare machine learning. Therefore, SHAP-based feature analysis provides a more credible and robust framework for assessing feature relevance in clinical predictive modeling.

REFERENCES

[1] D. I. Miguel, A. S. Solange N., I. Marcia, and Da Silva Suzana Alves, "Data Quality in health records: A literature review," *Iberian Conference on Information Systems and Technologies, CISTI*, Jun. 2021, doi: 10.23919/CISTI52073.2021.9476536.

[2] H. Ali, I. K. Niazi, B. K. Russell, C. Crofts, S. Madanian, and D. White, "Review of Time Domain Electronic Medical Record Taxonomies in the Application of Machine Learning," *Electronics 2023, Vol. 12, Page 554*, vol. 12, no. 3, p. 554, Jan. 2023, doi: 10.3390/ELECTRONICS12030554.

[3] F. Diaz-Garelli *et al.*, "What Oncologists Want: Identifying Challenges and Preferences on Diagnosis Data Entry to Reduce EHR-Induced Burden and Improve Clinical Data Quality," *JCO Clinical Cancer Informatics*, no. 5, pp. 527–540, Dec. 2021, doi: 10.1200/CCI.20.00174;JOURNAL:JOURNAL:CCI;WGROUP:STRING:PUBLICATION.

[4] M. K. Siam, M. J. Hossain Faruk, B. He, J. Q. Cheng, and H. Gu, "Multimodal Models in Healthcare: Methods, Challenges, and Future Directions for Enhanced Clinical Decision Support," *Information 2025, Vol. 16, Page 971*, vol. 16, no. 11, p. 971, Nov. 2025, doi: 10.3390/INFO16110971.

[5] Y. Yu, M. Li, L. Liu, Y. Li, and J. Wang, "Clinical big data and deep learning: Applications, challenges, and future outlooks," *Big Data Mining and Analytics*, vol. 2, no. 4, pp. 288–305, 2019, doi: 10.26599/BDMA.2019.9020007.

[6] J. Latif, C. Xiao, S. Tu, S. U. Rehman, A. Imran, and A. Bilal, "Implementation and Use of Disease Diagnosis Systems for Electronic Medical Records Based on Machine Learning: A Complete Review," *IEEE Access*, vol. 8, pp. 150489–150513, 2020, doi: 10.1109/ACCESS.2020.3016782.

[7] M. E. Hossain, A. Khan, M. A. Moni, and S. Uddin, "Use of Electronic Health Data for Disease Prediction: A Comprehensive Literature Review," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 18, no. 2, pp. 745–758, 2021, doi: 10.1109/TCBB.2019.2937862.

[8] S. Shurrab, A. Guerra-Manzanares, A. Magid, B. Piechowski-Jozwiak, S. F. Atashzar, and F. E. Shamout, "Multimodal Machine Learning for Stroke Prognosis and Diagnosis: A Systematic Review," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 11, pp. 6958–6973, 2024, doi: 10.1109/JBHI.2024.3448238.

[9] A. Moncada-Torres, M. C. van Maaren, M. P. Hendriks, S. Siesling, and G. Geleijnse, "Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival," *Scientific Reports 2021 11:1*, vol. 11, no. 1, pp. 6968-, Mar. 2021, doi: 10.1038/s41598-021-86327-7.

[10] I. Kolyshkina and S. Simoff, "Interpretability of Machine Learning Solutions in Public Healthcare: The CRISP-ML Approach," *Frontiers in Big Data*, vol. 4, no. May, 2021, doi: 10.3389/fdata.2021.660206.

[11] M. T. Lehto, T. Kauppila, H. Kautiainen, M. K. Laine, O. Rahkonen, and K. H. Pitkälä, "Symptomatic diagnoses in primary care: an observational cohort study," *BJGP Open*, vol. 8, no. 4, pp. 1–11, 2024, doi: 10.3399/BJGPO.2023.0234.

[12] P. R. Chinthalapelly and S. Gorle, "AI-Enhanced Data Integration Frameworks for Multimodal Clinical Research," *Journal of AI-Powered Medical Innovations (International online ISSN 3078-1930)*, vol. 1, no. 1, pp. 119–129, 2024, doi: 10.60087/japmi.vol01.issue01.p129.

[13] A. Kline *et al.*, "Multimodal machine learning in precision health: A scoping review," *npj Digital Medicine*, vol. 5, no. 1, pp. 1–14, 2022, doi: 10.1038/s41746-022-00712-8.

[14] A. CHAMMA, D. A. Engemann, and B. Thirion, "Statistically Valid Variable Importance Assessment through Conditional Permutations," *Advances in Neural Information Processing Systems*, vol. 36, pp. 67662–67685, Dec. 2023.

[15] I. Altaf, M. C.-J. of S. C. and Data, and undefined 2024, "Permuted Gini Importance–PaP Impurity Measurement for Tree-Based Models," *penerbit.uthm.edu.myI Altaf, MA ChachooJournal of Soft Computing and Data Mining, 2024•penerbit.uthm.edu.my*, vol. 5, no. 2, pp. 96–109, 2024, doi: 10.30880/jscdm.2024.05.02.008.

[16] K. Bladen, D. R. Cutler, K. Bladen, and D. R. Cutler, "Assessing agreement between permutation and dropout variable importance methods for regression and random forest models," *Electronic Research Archive 2024 7:4495*, vol. 32, no. 7, pp. 4495–4514, 2024, doi: 10.3934/ERA.2024203.

[17] T. Mehari *et al.*, "ECG Feature Importance Rankings: Cardiologists Vs. Algorithms," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 4, pp. 2014–2024, Apr. 2024, doi: 10.1109/JBHI.2024.3354301.

[18] Y. Takefuji, "Reevaluating feature importances in machine learning models for schizophrenia and bipolar disorder: The need for true associations," *Brain, Behavior, and Immunity*, vol. 124, pp. 123–124, Feb. 2025, doi: 10.1016/J.BBI.2024.11.036.

[19] C. Cordova *et al.*, "HER2 classification in breast cancer cells: A new explainable machine learning application for immunohistochemistry," *Oncology Letters*, vol. 25, no. 2, pp. 1–9, Feb. 2023, doi: 10.3892/OL.2022.13630/ABSTRACT.

[20] F. Wang, D. Wang, Y. Xu, H. Jiang, Y. Liu, and J. Zhang, "Potential of the Non-Contrast-Enhanced Chest CT Radiomics to Distinguish Molecular Subtypes of Breast Cancer: A Retrospective Study," *Frontiers in Oncology*, vol. 12, p. 848726, Mar. 2022, doi: 10.3389/FONC.2022.848726/BIBTEX.

[21] S. M. Lundberg, P. G. Allen, and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[22] J. Sun, C. K. Sun, Y. X. Tang, T. C. Liu, and C. J. Lu, "Application of SHAP for Explainable Machine Learning on Age-Based Subgrouping Mammography Questionnaire Data for Positive Mammography Prediction and Risk Factor Identification," *Healthcare 2023, Vol. 11, Page 2000*, vol. 11, no. 14, p. 2000, Jul. 2023, doi: 10.3390/HEALTHCARE11142000.

[23] N. Hettikankanamage, N. Shafiabady, F. Chatteur, R. M. X. Wu, F. Ud Din, and J. Zhou, "eXplainable Artificial Intelligence (XAI): A Systematic Review for Unveiling the Black Box Models and Their Relevance to Biomedical Imaging and Sensing," *Sensors*, vol. 25, no. 21, p. 6649, Nov. 2025, doi: 10.3390/S25216649/S1.

[24] I. D. Mienye *et al.*, "A survey of explainable artificial intelligence in healthcare: Concepts, applications, and challenges," *Informatics in Medicine Unlocked*, vol. 51, no. October, p. 101587, 2024, doi: 10.1016/j.imu.2024.101587.

[25] A. Saranya and R. Subhashini, "A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends," *Decision Analytics Journal*, vol. 7, no. February, p. 100230, 2023, doi: 10.1016/j.dajour.2023.100230.
[26] Z. Sadeghi *et al.*, "A review of Explainable Artificial Intelligence in healthcare," *Computers and Electrical Engineering*, vol. 118, no. PA, p. 109370, 2024, doi: 10.1016/j.compeleceng.2024.109370.
[27] A. M. Salih *et al.*, "A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME," *Advanced Intelligent Systems*, vol. 7, no. 1, p. 2400304, Jan. 2025, doi: 10.1002/AISY.202400304;WGROUP:STRING:PUBLICATION.
[28] A. M. Salih, I. B. Galazzo, Z. Raisi-Estabragh, S. E. Petersen, G. Menegaz, and P. Radeva, "Characterizing the Contribution of Dependent Features in XAI Methods," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 11, pp. 6466–6473, 2024, doi: 10.1109/JBHI.2024.3395289.
[29] S. Studer *et al.*, "Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology," *Machine Learning and Knowledge Extraction*, vol. 3, no. 2, pp. 392–413, 2021, doi: 10.3390/make3020020.
[30] N. Mumtazah, I. Waspada, K. N. Dinar Mutiara, and S. Adhy, "A Combination of Data Mining Methods for Disease Classification Using Patient-Perceived Symptoms from Medical Records," *Proceedings - International Conference on Informatics and Computational Sciences*, pp. 426–431, 2024, doi: 10.1109/ICICoS62600.2024.10636866.
[31] K. M. Chaitrashree, T. N. Sneha, S. R. Tanushree, G. R. Usha, and T. C. Pramod, "Unstructured Medical Text Classification using Machine Learning and Deep Learning Approaches," *2021 6th International Conference on Recent Trends on Electronics, Information, Communication and Technology, RTEICT 2021*, pp. 429–433, 2021, doi: 10.1109/RTEICT52294.2021.9573667.
[32] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.
[33] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 63, no. 1, pp. 3–42, 2006, doi: 10.1007/s10994-006-6226-1.
[34] T. Chen and C. Guestrin, "XGBoost," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: ACM, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.
[35] G. Ke *et al.*, "LightGBM: A highly efficient gradient boosting decision tree," *Advances in Neural Information Processing Systems*, vol. 2017-Decem, no. Nips, pp. 3147–3155, 2017.
[36] V. N. Vapnik, "An overview of statistical learning theory," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 988–999, 1999, doi: 10.1109/72.788640.
[37] A. B. Mahammad and R. Kumar, "Design a Linear Classification model with Support Vector Machine Algorithm on Autoimmune Disease data," *Proceedings of 3rd International Conference on Intelligent Engineering and Management, ICIEM 2022*, pp. 164–169, 2022, doi: 10.1109/ICIEM54221.2022.9853182.
[38] H. Han, M. Huang, Y. Zhang, and J. Liu, "Decision support system for medical diagnosis utilizing imbalanced clinical data," *Applied Sciences (Switzerland)*, vol. 8, no. 9, 2018, doi: 10.3390/app8091597.
[39] M. Salmi, D. Atif, D. Oliva, A. Abraham, and S. Ventura, *Handling imbalanced medical datasets: review of a decade of research*, vol. 57, no. 10. Springer Netherlands, 2024. doi: 10.1007/s10462-024-10884-2.
[40] M. H. Rahman, M. N. Khan, S. Das, and B. Uddin, "Explainable AI Framework for Precision Public Health in Metabolic Disorders: A Federated, Multi-Modal Predictive Modelling Approach for Early Detection and Intervention of Type 2 Diabetes," *The Eastasouth Journal of Information System and Computer Science*, vol. 3, no. 02, pp. 164–178, 2025, doi: 10.58812/esiscs.v3i02.759.