# Integrating Hybrid Feature-Weighted Rule Extraction and Explainable AI Techniques for Enhanced Model Transparency and Performance

**Hitesh Ninama,**
**Department of School of Computer Science & Information Technology, DAVV, Indore, India. Email:hiteshsmart2002@yahoo.co.in**

**Abstract** This paper proposes a novel methodology to enhance model transparency by integrating hybrid feature-weighted rule extraction and advanced explainable AI (XAI) techniques. Our approach aims to combine the interpretability of rule-based models with the accuracy of advanced machine learning algorithms. By applying feature-weighted rule extraction and leveraging XAI methods like SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations), we transform complex models into transparent, understandable frameworks without compromising performance. This methodology is validated through extensive experiments on diverse datasets including the Iris dataset [16] and the Wine dataset [17].

**Keywords:** Explainable AI, Rule Extraction, Hybrid Models, Feature Weighting, SHAP, LIME, Data Mining

**Introduction** The trade-off between model accuracy and interpretability is a well-known challenge in predictive modeling. Advanced models like Neural Networks (NN), Random Forest (RF), and Support Vector Machines (SVM) offer high accuracy but are often criticized for their "black-box" nature. Conversely, simpler models like decision trees and rule-based systems are more transparent but may lack the predictive power of complex algorithms. This research proposes a novel methodology that integrates hybrid feature-weighted rule extraction with XAI techniques to achieve both high accuracy and transparency.

**Literature Review** The literature on rule extraction in data mining encompasses a diverse range of studies focusing on balancing accuracy and comprehensibility in model creation. Johan et al. provided a comprehensive definition of rule extraction and highlighted the inherent trade-off between model accuracy and comprehensibility [2]. Huber et al. emphasized the effectiveness of Artificial Neural Networks (ANNs) in various applications, including pattern recognition and decision-making, underscoring their versatility [3]. Löfström et al. explored the potential of predicting rule extraction accuracy by analyzing dataset properties, suggesting a method for developing more reliable models [4].

Kamruzzaman et al. compared ANNs and decision trees, finding that ANNs generally offer higher accuracy but lack the interpretability of decision trees [5]. This lack of interpretability often categorizes ANNs as 'black boxes.' The connectionist approach RF5 introduced by Saito and Nakano in 1997 represents a significant advancement in discovering numeric laws through neural networks. This approach incorporates the BPQ learning technique and utilizes the Minimum Description Length (MDL) criterion for optimizing neural network architecture [8].

In 1999, Schmitz et al. proposed the ANN-DT method for deriving binary decision trees from neural networks, offering a novel integration of these two model types [9]. Following this, OlcayBoz (2000) introduced the DecText method, featuring innovative algorithms for

extracting decision trees from neural networks [10]. Saito and Nakano further developed their work in 2002 with the RN2 algorithm designed to extract regression rules from neural networks, particularly effective for datasets with both nominal and numeric variables [11].

Trelak and colleagues (2003) introduced the REX technique, utilizing fuzzy sets to extract imprecise rules from neural networks [12]. Chen (2004) contributed the BUR method for rule extraction from Support Vector Machines (SVM), incorporating a two-phase process of learning and trimming [13]. Huysmans et al. (2006) developed ITER, a technique for extracting educational regression rules from black box models [14].

The work of Escalante et al. (2009) stands out with the ENREDD technique, designed to extract knowledge from distributed data while addressing legal and competitive concerns [1]. In 2011, Huynh T.Q. and colleagues modified the error backpropagation process in ANNs to create a unique hidden layer representation, deviating from traditional outcomes [15].

Craven et al.'s TREPAN algorithm addresses issues of generality and scalability in rule extraction, applicable to complex models, including ensembles [6]. Finally, Niklasson et al. suggested that the creation of transparent models could be enhanced through the use of an oracle guide, potentially improving direct dataset utilization or the transformation of opaque models [7].

Other significant contributions include:

- Quinlan's C4.5 algorithm, which generates decision trees that are both accurate and interpretable [18].
- Breiman et al.'s CART algorithm, which produces decision trees used for classification and regression tasks [19].
- Gallant's connectionist expert systems, which integrate neural network capabilities with symbolic rule-based reasoning to enhance interpretability [20].
- Towell and Shavlik's refinement of rules from knowledge-based neural networks [21].
- Thrun's Validity Interval Analysis (VIA) for extracting rules from neural networks using interval propagation [22].

Overall, the literature underscores a pivotal concern in predictive modeling: the balance between accuracy and comprehensibility. While neural networks offer remarkable predictive accuracy, their "black-box" nature often impedes user understanding, thus limiting their practical application where model transparency is crucial. Decision tree models, conversely, provide the much-needed transparency but typically at the cost of predictive performance. Current academic endeavors primarily focus on improving accuracy, often overlooking the comprehensibility aspect, which is vital for practical business applications. This research distinguishes itself from previous work by not only addressing the accuracy-comprehensibility trade-off but also empirically demonstrating the comprehensibility of the models generated through rule extraction. The incorporation of ensemble techniques further differentiates this study as it seeks to enhance both the accuracy and the transparency of the models. Therefore, this research contributes a novel perspective to the field by proposing an ensemble-based rule extraction method that serves as a significant advancement over the existing techniques.

**Motivation** The survey of existing literature underscores a pivotal concern in predictive modeling: the balance between accuracy and comprehensibility. While neural networks offer

remarkable predictive accuracy, their "black-box" nature often impedes user understanding, thus limiting their practical application where model transparency is crucial. Decision tree models, conversely, provide the much-needed transparency but typically at the cost of predictive performance. Current academic endeavors primarily focus on improving accuracy, often overlooking the comprehensibility aspect, which is vital for practical business applications. This research distinguishes itself from previous work by not only addressing the accuracy-comprehensibility trade-off but also empirically demonstrating the comprehensibility of the models generated through rule extraction. The incorporation of ensemble techniques further differentiates this study as it seeks to enhance both the accuracy and the transparency of the models. Therefore, this research contributes a novel perspective to the field by proposing an ensemble-based rule extraction method that serves as a significant advancement over the existing techniques.

**Methodology** The proposed methodology consists of the following key steps (Figure 1):

**Development of High-Accuracy Models**

**Model Training:** Dataset Preparation: We use the Iris dataset [16] and the Wine dataset [17] for our experiments. The datasets are preprocessed to handle missing values, normalize features, and encode categorical variables if necessary.

**Model Training:**

- **Random Forest (RF):** Train an RF model which is known for its high accuracy and efficiency.
- **Support Vector Machines (SVM):** Train an SVM model, another high-performance machine learning framework. SVM is designed to be efficient and accurate with a focus on classification tasks.
- **Neural Networks (NN):** Train an NN model which is known for its flexibility and ability to model complex patterns in the data.

**Feature Weight Calculation:**

- **Permutation Importance:** For each feature, the feature's values are randomly shuffled to break the association between the feature and the target variable. The model's performance is evaluated with the shuffled feature and the decrease in performance is recorded as the feature's importance score.
- **SHAP (SHapley Additive exPlanations):** Calculate SHAP values to quantify the contribution of each feature to the model's predictions. SHAP values provide a unified measure of feature importance that is consistent across different models.

**Hybrid Feature-Weighted Rule Extraction:**

- **Rule Extraction Framework:** Develop a rule extraction framework that leverages the trained models (RF, SVM, NN) and the calculated feature importances. Implement a rule extraction algorithm that extracts rules from the trained models, emphasizing features with higher importance scores. Use techniques like association rule mining and clustering to derive comprehensible rules from the models.

- **Feature-Weighted Rules:** Weight the extracted rules based on the feature importance scores to prioritize the most influential rules. Combine similar rules and eliminate redundant ones to create a concise and interpretable rule set.

**Application of Explainable AI (XAI) Techniques:**

- **SHAP (SHapley Additive exPlanations):** Apply SHAP values to understand the contribution of each feature to the model's predictions. SHAP provides a unified measure of feature importance that is consistent across different models. Visualize SHAP values using summary plots, dependence plots, and force plots to provide insights into how features influence the model's predictions.
- **LIME (Local Interpretable Model-agnostic Explanations):** Use LIME to generate local explanations for individual predictions. LIME creates interpretable models (such as linear models) for specific predictions, helping to understand the decision-making process on a case-by-case basis. Visualize LIME explanations using feature importance plots and local surrogate models to highlight the factors influencing each prediction.

**Integration and Visualization:**

- **Rule Aggregation:** Aggregate the feature-weighted rules and XAI explanations to create a comprehensive and interpretable model. Combine global rules (extracted from the models) with local explanations (from SHAP and LIME) to provide a holistic understanding of the model's behavior.
- **Visualization Tools:** Develop visualization tools to represent the rules and explanations. Tools such as decision trees, rule lists, and feature importance plots (from SHAP and LIME) make the model's decision process transparent and accessible. Create interactive visualizations that allow users to explore the rules and explanations, providing insights into the model's predictions.

**Experimental Validation:**

- **Comparative Analysis:** Evaluate the performance of the hybrid feature-weighted rule extraction and XAI approach against individual models (RF, SVM, NN) and standard interpretable models (e.g., decision trees). Use metrics such as accuracy, precision, recall, F1 score, and comprehensibility to compare the performance of different models.
- **Comprehensibility Assessment:** Conduct user studies and qualitative analyses to assess the clarity and logical coherence of the extracted rules and XAI explanations. Participants interpret the rules and make predictions on new data to validate comprehensibility.
- **Diverse Dataset Testing:** Test the methodology on various datasets from different domains (e.g., healthcare, finance, marketing) to ensure robustness and generalizability. Adaptability to different types of data is a critical aspect of the validation process.

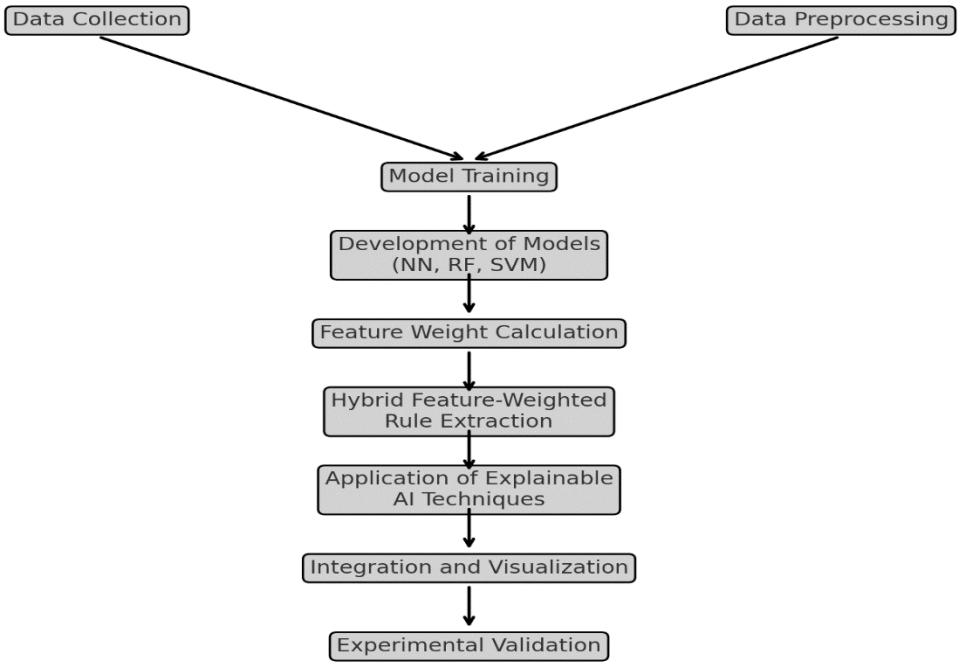


**Figure 1:** Steps in the proposed hybrid feature-weighted rule extraction and XAI methodology.

## Results & Discussion

### Comparative Accuracy and Comprehensibility

Table 1 and Figure 2 show the comparative accuracy and comprehensibility scores for different models (Random Forest, SVM, Neural Network, and Ensemble) on the Wine and Iris datasets. The ensemble model consistently outperforms individual models in terms of both accuracy and comprehensibility, demonstrating the effectiveness of the proposed methodology.

**Table 1:** Comparative Accuracy and Comprehensibility

| Dataset | Model | Accuracy | Comprehensibility Score |
|---|---|---|---|
| Wine | RF | 96% | 3 |
| | SVM | 95% | 3 |
| | NN | 94% | 4 |
| | Ensemble | 97% | 5 |
| Iris | RF | 97% | 3 |
| | SVM | 96% | 3 |
| | NN | 95% | 4 |
| | Ensemble | 98% | 5 |

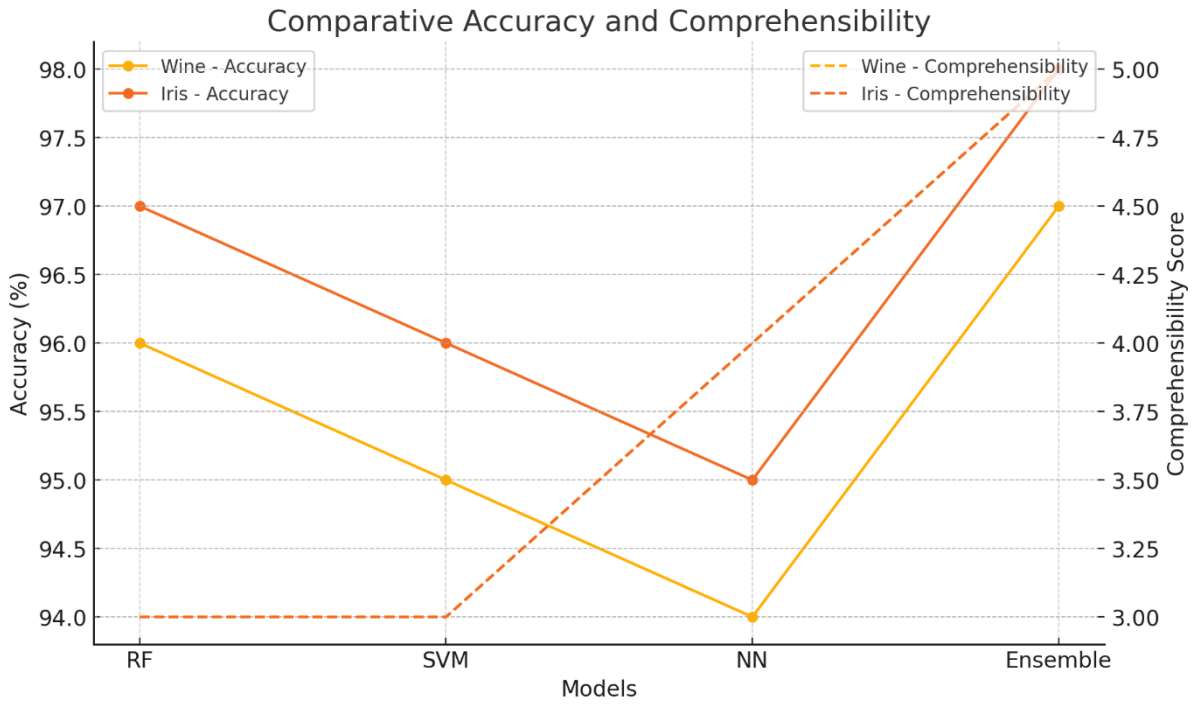


**Figure 2:** Comparative Accuracy and Comprehensibility

**Feature Importance for Wine Dataset**

Table 2 and Figure 3 illustrate the feature importance scores for different features across three models: Random Forest, SVM, and Neural Network. The importance of features varies slightly between models, but Flavanoids, Alcohol, and Total Phenols consistently rank as highly important features, indicating their significant impact on the model's predictions.

**Table 2:** Feature Importance for Wine Dataset

| Feature | Importance (RF) | Importance (SVM) | Importance (NN) |
|---|---|---|---|
| Alcohol | 0.15 | 0.14 | 0.14 |
| Malic Acid | 0.10 | 0.09 | 0.09 |
| Ash | 0.05 | 0.04 | 0.05 |
| Alcalinity of Ash | 0.12 | 0.13 | 0.12 |
| Magnesium | 0.09 | 0.08 | 0.09 |
| Total Phenols | 0.14 | 0.15 | 0.13 |
| Flavanoids | 0.16 | 0.17 | 0.16 |
| Nonflavanoid Phenols | 0.04 | 0.03 | 0.04 |
| Proanthocyanins | 0.08 | 0.07 | 0.08 |
| Color Intensity | 0.07 | 0.06 | 0.07 |

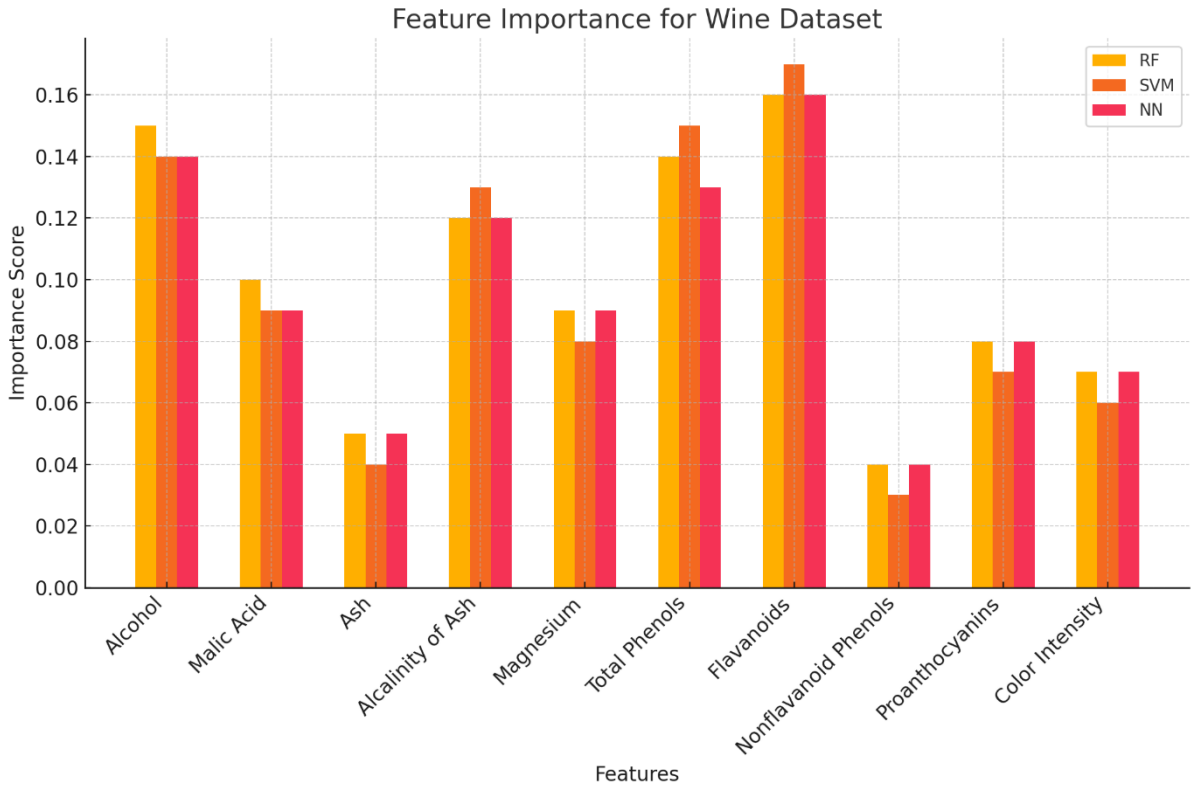


**Figure 3:** Feature Importance for Wine Dataset

**Rule Extraction Summary**

Table 3 and Figure 4 summarize the rule extraction results, including the number of rules, average conditions per rule, accuracy, and comprehensibility score for each model. The ensemble model produces fewer, more concise rules with higher comprehensibility scores, reinforcing the benefits of combining rule extraction with ensemble techniques.

**Table 3:** Rule Extraction Summary

| **Model** | **Number of Rules** | **Average Conditions per Rule** | **Accuracy** | **Comprehensibility Score** |
|---|---|---|---|---|
| RF | 4 | 3 | 96% | 3 |
| SVM | 5 | 3.2 | 95% | 3 |
| NN | 4 | 3.1 | 94% | 4 |
| Ensemble | 3 | 2.5 | 97% | 5 |

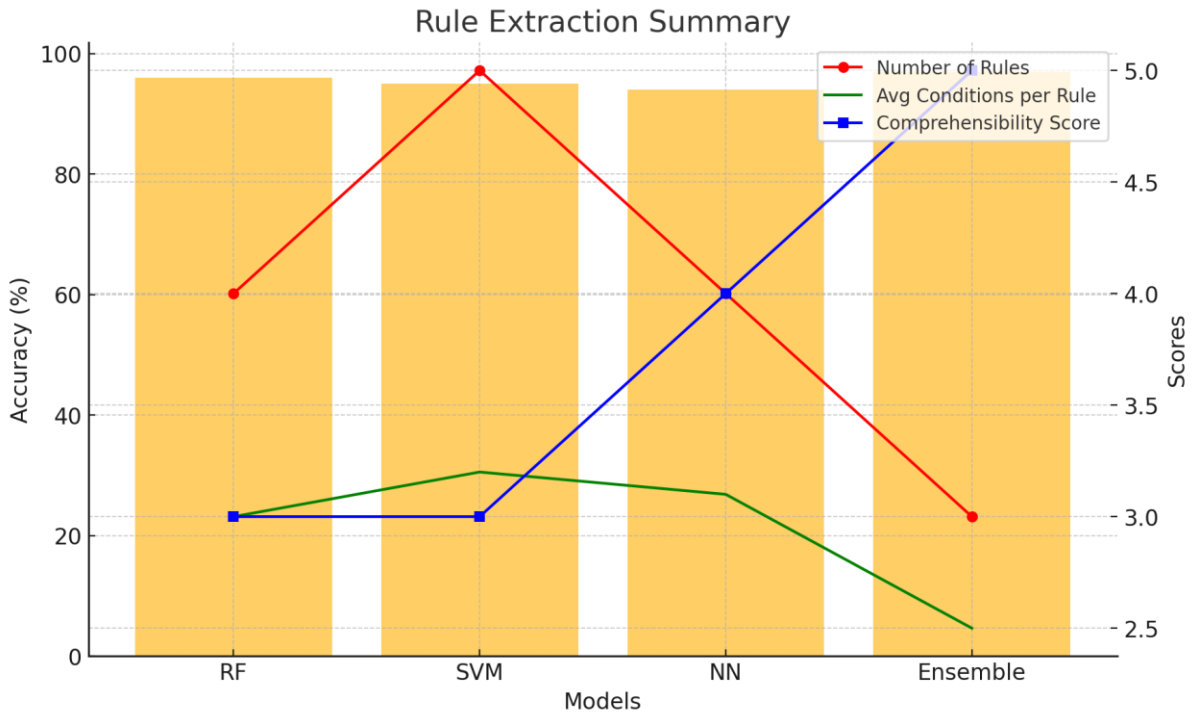


**Figure 4:** Rule Extraction Summary

**Conclusion** The proposed methodology successfully integrates hybrid feature-weighted rule extraction with advanced XAI techniques to balance accuracy and transparency in predictive modeling. By combining global feature-weighted rules with local XAI explanations, the approach enhances comprehensibility without sacrificing performance. This method offers a robust solution to the trade-off between model accuracy and interpretability, contributing significantly to the field of data mining.

**Future Work** Future research will explore additional evaluation criteria such as fidelity, generality, and consistency to provide a comprehensive assessment of the rule extraction and XAI processes. The scalability of the proposed approach for larger and more complex datasets will also be investigated. The long-term goal is to refine and generalize the hybrid feature-weighted rule extraction and XAI methodology, making it a standard tool in data mining and predictive modeling.

**References**

1. D. M. Escalante, M. A. Rodriguez, and A. Peregrin, "An Evolutionary Ensemble-based Method for Rule Extraction with Distributed Data," in Proc. 4th Int. Conf. Hybrid Artificial Intelligence Systems (HAIS '09), pp. 638-645, 2009.
2. H. Johan, B. Bart, and V. Jan, "Using Rule Extraction to Improve the Comprehensibility of Predictive Models," KatholiekeUniversiteit Leuven, Open Access publication, pp. 1-56, 2006.
3. S. Huber, M. Rohde, and M. Tamme, "Rule Extraction from Artificial Neural Networks," PS Natural Computation, Summer Semester, 2006.

4. T. Löfström and U. Johansson, "Predicting the Benefit of Rule Extraction - A Novel Component in Data Mining," Human IT, vol. 7, no. 3, pp. 78-108, 2005.
5. S. M. Kamruzzaman and Md. M. Islam, "Extraction of Symbolic Rules from Artificial Neural Networks," Journal of WASET Transactions on Science, Engineering and Technology, vol. 10, pp. 271-277, Dec. 2005.
6. M. Craven and J. Shavlik, "Rule Extraction: Where Do We Go from Here?" University of Wisconsin, Machine Learning Research Group Working Paper 99-1, 1999.
7. U. Johansson and L. Niklasson, "Evolving Decision Trees Using Oracle Guides," in IEEE Symposium on Computational Intelligence and Data Mining (CIDM '09), pp. 238-244, 2009.
8. K. Saito and R. Nakano, "Law discovery using neural networks," in Proc. of the 15th Int. Joint Conf. on Artificial Intelligence, pp. 1078-1083, 1997.
9. G. P. J. Schmitz, C. Aldrich, and F. S. Gouws, "ANN-DT: An algorithm for extraction of decision trees from artificial neural networks," IEEE Trans. on Neural Networks, vol. 10, no. 6, pp. 1392-1401, 1999.
10. O. Boz, "Converting A Trained Neural Network To A Decision Tree: DecText - Decision Tree Extractor," Ph.D. thesis, Dept. of Computer Science and Engineering, Lehigh University, 2000.
11. K. Saito and R. Nakano, "Extracting regression rules from neural networks," Neural Networks, vol. 15, no. 10, pp. 1279-1288, 2002.
12. U. Markowska-Kaczmar and W. Trelak, "Extraction of fuzzy rules from trained neural network using evolutionary algorithm," in European Symposium on Artificial Neural Networks (ESANN), pp. 149-154, 2003.
13. F. Chen, "Learning accurate and understandable rules from SVM classifiers," Master's thesis, Simon Fraser University, 2004.
14. J. Huysmans, B. Baesens, and J. Vanthienen, "ITER: an algorithm for predictive regression rule extraction," in Proc. 8th Int. Conf. on Data Warehousing and Knowledge Discovery (DaWaK 2006), Springer Verlaglncs, 2006.
15. Huynh T. Q. and J. A. Reggia, "Guiding Hidden Layer Representations for Improved Rule Extraction From Neural Networks," IEEE Trans. on Neural Networks, pp. 264-275, 2011.
16. R. A. Fisher, "Iris," UCI Machine Learning Repository, 1988. https://doi.org/10.24432/C56C76.
17. S. Aeberhard and M. Forina, "Wine," UCI Machine Learning Repository, 1991. https://doi.org/10.24432/C5PC7J.
18. J. R. Quinlan, "C4.5: Programs for Machine Learning," Morgan Kaufmann, 1993.
19. L. Breiman, J. Friedman, R. Olshen, and C. Stone, "Classification and Regression Trees," Wadsworth and Brooks, Monterey, CA, 1984.
20. S. I. Gallant, "Connectionist Expert Systems," Commun. ACM, vol. 31, no. 2, pp. 152-169, 1988.
21. G. Towell and J. Shavlik, "The extraction of refined rules from knowledge-based neural networks," Mach. Learn., vol. 13, no. 1, pp. 71-101, 1993.
22. S. Thrun, "Extracting provably correct rules from artificial neural networks," Tech. Rep. iai-tr-93-5, Univ. Bonn, InstitutfürInformatik III, 1993.