WCRG Global Research
WORLD COMET RESEARCH GROUP

International Journal of Applied Data Science & Modern Computing
Vol. 6, No. 1, 2023
Doi: 10.67228/3142788X/IJMLPA-2023PI9W2R
PP. 01-12

Original Article

Interpretable Machine Learning Models for Critical Decision Systems

Amanda Davis

Business Development Manager, Nexus Solutions, USA.

Received: 04-02-2023 Revised: 04-20-2023 Accepted: 05-02-2023 Published: 05-05-2023

ABSTRACT

Machine learning (ML) systems are finding more and more applications in high-reliability decision-making setting including medical diagnosis, risk estimation in the finance sector, driverless transport, law enforcement, and fault management in industries. Their lack of transparency makes complex black-box models, especially deep neural networks and ensemble learning methods, highly problematic in safety-critical and high-stakes application areas despite having shown impressive predictive accuracy. The regulatory requirement, ethical concerns, accountability and trust among users enforce that the decisions made by ML systems must be interpretable, clarifiable and verifiable. That caused the increased attention to the area of interpretable machine learning (IML), which is supposed to reconcile predictive score and interpretable reasoning available to humans. This paper is the systematic and complete study of interpretable machine learning models of critical decision systems. We start by examining the conceptual basis behind interpretability and its significance in high risk applications. An elaborate literature review classifies the currently existing interpretability methods as intrinsic interpretability methods and post-hoc explanation methods and their strong and weak points and the appropriateness to critical systems. The suggested methodology describes a systematic approach to the selection, design and validation of interpretable ML models within real-life conditions of uncertainties of data, bias and regulatory standards. Mathematically stated representative interpretable models such as linear models, decision trees, rule-based systems and attention mechanisms are given to provide formal grounding. Examples of experimental findings of representative domains of application are presented to illustrate the trade-offs in interpretability and performance. The discussion highlights levels of interpretability, robustness, and fairness measures, and predictive accuracy. Lastly, the paper draws a conclusion and gives important opinions and future research directions with the aim of achieving credible and open machine learning systems in life and death situations.

KEYWORDS

Interpretable Machine Learning, Explainable AI, Critical Decision Systems, Model Transparency, Trustworthy AI, Fairness, Accountability, Safety-Critical Applications.

Copyright © WCRG The Author(s) This article is distributed under a
Creative Commons Attribution-Share A like 4.0 International License.

WCRG
Global Research
ISSN: 3142-788X
PP: 01-12.

1. INTRODUCTION

1.1. Background

As a result of the rapid development of machine learning technologies, the process of decision-making has been radically transformed in a vast scope of essential industries, such as the healthcare industry, the financial sector, autonomous movement of transport, governance, and law enforcement. In such areas, machine learning models are the growingly preferred assistants or completely automatic motivational choices, comprising medical evaluation, credit danger, frauds, road guidance, and government policy evaluation. Although these systems introduce significant change in efficiency and prediction abilities, consequences of an erroneous, prejudiced, or misconceived prediction may be significant, and may lead to the loss of human life, economic disturbance, legal claims or social injustice. With the increasing interdependence on the automated decision systems, there is the need to make sure that they are reliable, just, and they adhere to the societal values. However, even though they proved to be effective enough, a significant number of state-of-the-art machine learning models operate as black boxes, offering forecasts without any obvious information of how those forecasts are reached. Deep neural networks, gradient-boosted ensemble algorithms, and other complex architectures usually tend to compromise transparency in advance of prediction. Such lack of interpretability presents major problems in critical applications where decisions have to be justified, audited and relied by human stakeholders. Taking clinicians as an example, they need to be provided with comprehensible explanations to confirm their diagnostic recommendations, financial regulators use transparency in lending and risk assessment decisions, and developers of autonomous systems need to demonstrate explainable behavior to meet safety certification and liability standards. Interpretable machine learning has become an essential reaction to these issues since it pays attention to how models can be constructed their decision-making can be viewed as understandable or has contribution to clarifying how complex models behave into could be provided. Interpretability is not only a technical issue, but also a general principle of building trust and accountability and ethical responsibility in artificial intelligence systems. Making automated decisions explainable through interpretable machine learning helps implement AI technology in critical decision systems, where transparency is needed, to enable individuals to challenge and verify it.

1.2. Importance of Interpretability in Critical Systems

Fig 1 - Importance of Interpretability in Critical Systems

ISSN: 3142-788X
PP: 01-12.

1.2.1. Trust and User Confidence

Interpretability the services that establish the trust between machine learning systems and human stakeholders in environments that demand high reliability rely on interpretability. The more users like the clinicians, financial analysts or system operators understand the route that a model took to reach a decision, the more they will trust in its suggestions and make proper use of it. The explicit decision logic can help the user test the results of a system with the knowledge of the domain and help in detecting the possible error and building trust in the reliability of a system. Conversely, when Opaque models are used one may end up with a skeptical and resistant reaction and adoption restraint although it may be highly predictive.

1.2.2. Accountability and Explainability

In critical systems, decision-making should be unambiguable so as to hold responsibilities as out of the consequences of decision making, outcomes have legal, ethical or social consequences. Interpretability also enables businesses to identify decisions based on particular inputs and well-defined reasoning processes, and thus, actions made by automated systems can be justified. It is especially significant in the situations of conflicts, audit, or insurance evaluations, where the stakeholders have to present the reasons why a specific choice was taken and who is to be an author of this choice.

1.2.3. Regulatory and Legal Compliance

A lot of regulated business such as health care, finance and government administration have very stringent demand on the transparency and equity in making decisions through automated means. Interpretability aids organisations in adhering to regulation that requires them to give an explanation of decisions that include individuals like loan issuance or a medical advice. Models that contain simple and comprehensible logic are simpler to audit and certify and minimize chances of breaking the rules and legal disputes.

1.2.4. Safety and Risk Mitigation

Interpretability in safety-critical systems is important in the identification and reduction of risks. Knowledge of model behavior allows engineers and subject matter experts to identify unsafe decision patterns, failures, or bias in their models prior to deployment. Interpretable models are useful during operation, which can accelerate the diagnosis and correction of errors, improving system robustness and minimize the chances of catastrophic failures.

1.2.5. Ethical and Fair Decision-Making

The interpretability is crucial in the aspect of ensuring that machine learning is ethically utilized in critical systems. Transparent models enable the stakeholders to check the impact of biased or irrelevant features in decisions to promote fairness and inclusivity. Interpretability, by allowing the examination of the logic behind specific decisions, can be used to align AI systems with ethical standards and societal expectations, which solidifies responsible and human-centered AI deployment.

ISSN: 3142-788X
PP: 01-12.

1.3. Challenges in Interpretable Machine Learning

Even though interpretability in machine learning is increasingly becoming important in critical decision systems, a number of fundamental issues associated with it still restrict its prevalence and standardization in critical environments. The perceived complexity-interpretability of model trade-off is among the most obvious pitfalls. Deep neural networks and ensemble approaches are highly expressive models with the ability to capture complex nonlinear relationships and can sometimes do better in predictive performance. Nevertheless, they have complex interior representations that are very hard to explain and comprehend. By comparison, more basic models like linear classifiers or shallow decision trees provide transparency and can not necessarily capture the complex patterns of high dimensional (or unstructured) data. A tradeoff between predictive accuracy and interpretability is thus one of the major design challenges, especially when both performance and transparency are required. The other important obstacle is the subjective and contextual interpretation of interpretability. What is a meaningful explanation will differ significantly among stakeholders, such as domain experts, end users, regulators and developers. As an example, a clinician might need causal or clinically based explanations, a regulator might pay more attention to fairness and compliance and a system engineer might be more preoccupied with debugging and validation. All these multiplicity of expectations render it hard to formulate any general concept of interpretability that would meet every criterion. This leads to the interpretation solutions being, in many cases, application-specific, which adds more complexity to the development. Also, quantification of interpretability is a research issue. In comparison with predictive accuracy, which can be quantified on them with a set of well-defined statistical measures, interpretability entails human cognitive dimensions including comprehensibility, effort of reasoning, and trust, which are in and of themselves challenging to formalize and quantify in an objective way. Current proxy measures, e.g. model sparsity or rule length, merely partially reflect human comprehension and are unlikely to be practical regarding actual usability. To overcome these problems, it is necessary to conduct interdisciplinary studies incorporating machine learning, human-computer interaction, and cognitive science to design robust and human-friendly interpretability frameworks.

2. LITERATURE SURVEY

2.1. Intrinsically Interpretable Models

The intrinsically interpretable models are designed in a manner that the process of making decisions is transparent, and human users can move forward and see the impact of the inputs to outputs assessed directly. The literature has extensively mentioned the classical methods of linear regression, logistic regression, decision trees, and rule-based classifiers due to their straightforward nature and simplicity. Linear and logistic models provide the predictions in terms of the weighted linear combinations of input features, and this allows the easy interpretation of the input features significance and directional impact. Instead, decision trees represent decision logic as hierarchic rules, with each internal node representing a feature-based condition and each root-leaf path representing a decipherable decision rule. Rule-based classifiers are additionally more transparent with the explicit use of the if-then statements that are a reflection of the human reasoning. In spite of these benefits, intrinsic models have in most cases been limited when it comes to complex, nonlinear,

ISSN: 3142-788X
PP: 01-12.

or high-dimensional data typically used in practice. Recent studies have suggested improved interpretable models to overcome this like sparse linear models, shallow or constrained decision trees and generalized additive models (GAMs) which would allow trade off predictive expressiveness and interpretability. These advances reflect on continued attempts of expanding the area of applicability of intrinsic models without necessarily losing their appropriateness to high-stakes decision situations.

2.2. Post-hoc Explain ability Techniques

Research directions Post-hoc explain ability methods have become a popular stream of research to explain complex black-box models, such as deep neural networks and ensemble methods, without altering their internal architecture. Such methods will be especially appealing in case of high predictive accuracy needs, but the transparency of the model is necessary as well. The attainment of feature attribution methodology measures the impact of each input variable in prediction by a model, and therefore, allows users to know the influential variables. Local surrogate models are used to predict the behavior of a black box model in a local area around an instance, with localised human understandable explanations. Counterfactual explanations concentrate on determining the smallest changes to the input features that could change the model output and thus provide practical suggestions to the decision-makers. These methods have been applied heavily in various fields although the literature has observed significant issues concerning the reliability of the methods. Faithfulness, stability and consistency concerns still exist, and the explanations can be different using similar inputs or far too dependent on the explainer selected (not the model beneath it). The researchers, therefore, point to the necessity of a strict validation of post-hoc explanations especially in systems of high safety with false interpretation as having dire outcomes.

2.3. Interpretability in Safety-Critical Domains

Interpretability is an important concern in safety-critical environments where model choices refer to the human life in question, financial stability, or community safety. Interpretable machine learning models have demonstrated better clinician trust, diagnostic reasoning, and clinical decision-making in healthcare applications, because they can provide clear reasons why a prediction was made. Likewise, in the financial industry, interpretable credit scoring and risk assessment models are needed to understand the fairness of credit, accountability and adherence of the regulatory frameworks, which also facilitate the availability of traceability and validation of decisions by the auditors. Interpretability in autonomous and cyber-physical systems can help verify control policies, troubleshoot system failures, and also increase the safety of the system as a whole. It is always stated in the literature that transparent models enhance stakeholder confidence and ease certification and validation procedures. Consequently, interpretability is no longer considered a nice-to-have feature, but one of the preconditions of the implementation of machine learning systems in high-risk settings. The current research remains active in the directions of encompassing the concept of interpretability in the model development lifecycle to provide the view that the concepts of safety, reliability, and explainability can be developed simultaneously.

ISSN: 3142-788X
PP: 01-12.

3. METHODOLOGY

3.1. Proposed Framework for Interpretable Decision Systems

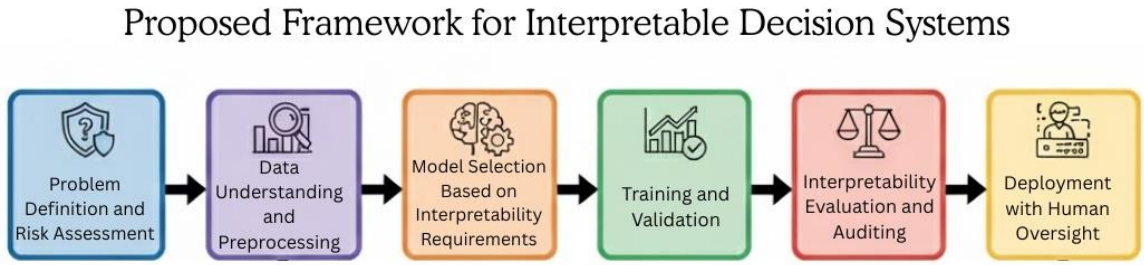

Fig 2 - Proposed Framework for Interpretable Decision Systems

3.1.1. Problem Definition and Risk Assessment

The initial step of the suggested model is the clear definition of the decision problem and the evaluation of the negative effects of the large-scale inaccurate or unclear model results. This step determines the area of application, the aim of decision, stakeholders, and possible unpleasant results of errors or bias. In controlled and safety-sensitive scenarios, e.g., in healthcare or finance, risk assessment involves considering the consequences of the failures of the models, their severity, the probability, and ethical aspect. This step informs how much interpretability is necessary and what type of explanation should be applied by classifying decisions according to their impact and regulation restrictions and so whether intrinsic or post-hoc methods of explanation can be used appropriately.

3.1.2. Data Understanding and Preprocessing

Data interpretation and preprocessing is a pivotal cornerstone of decision systems that can be interpreted. This step will consist of data sources analysis, features distributions, missing values, and possible biases that can affect model behavior. Feature engineering is done in an interpretable way, preferring variables that describe the domain to the use of complicated transformations. The methods of preprocessing, including normalization, discretization, and feature selection are used to improve the performance of the model and its transparency. Good data assumption and preprocessing processes are properly documented and can be traced back and help in auditability at subsequent levels.

3.1.3. Model Selection Based on Interpretability Requirements

The interpretability requirements that were observed during risk assessment are what dictate the model selection. In case of high-risk decisions, intrinsically interpretable models are selected (that is, the linear model, decision trees, or generalized additive model) since they have a transparent structure. In the situations, when the superior predictive accuracy in terms of the post-hoc explainability is necessary, sophisticated models can be employed. The stage is associated with judging the trade-offs between accuracy, interpretability, and computational complexity to ensure that the selected model is in accordance with the technical requirements as well as the expectations of the stakeholders.

ISSN: 3142-788X
PP: 01-12.

3.1.4. Training and Validation

Models are prepared in the training and validation stage which trains models with representative datasets whilst evaluating them with the required performance measures. The generalization and robustness are evaluated using cross-validation techniques, and interpretability-aware regularization is to be used to avoid overfitting and the decrease in complexity of models. Validation does not only mean accuracy but fairness, stability and consistency among the various subsets of data. Such a holistic assessment will make the model to work reliably and still have a comprehensible decision logic.

3.1.5. Interpretability Evaluation and Auditing

The domain knowledge and the human understanding Interpretability evaluation and auditing concentrate on generating systematic evaluation of how well model explanations are relevant to domain knowledge. This involves the examination of feature significance, choice directions, as well as, the reaction consistency on similar inputs. Auditing Commonly finding bias, spuriousness correlation, and drift over time Auditing procedures include checking bias, spuriousness correlation and drift over time. Human specialists can be engaged in order to certify answers and make them practical and reliable. This is a phase of instilling trust in the system especially in controlled or high stakes settings.

3.1.6. Deployment with Human Oversight

The last step entails the implementation of the interpretable decision system along with the systems of human control and intervention. The outputs and explanation of the models are provided in an easy-to-read format to be able to make informed decisions. The human-in-the-loop frameworks enable domain experts to second check, override or give feedback on model choices, and thus continue to learn and improve. Continued observation is done to identify the performance deterioration or changes in data distribution, to maintain reliability in the long term, accountability, and compliance.

3.2. Mathematical Formulation of Interpretable Models

Mathematical formulations describing the effects of input features on the predicted outcome are a defining feature of interpretable machine learning models, which provide transparency, as well as allow analytical reasoning. Linear models represent a commonly used type of interpretable models, in which the prediction is represented as an unbiased linear combination of weighted input features and a bias term. In this representation, a constant offset is added to the remittance of the total inputs of the contribution made by every component input history of the input feature to the predicted output. Every single coefficient has a direct representation of how strong and in what direction the relationship between a specific feature with the model prediction is, which gives the stakeholders a clear picture of the variables that influence the decision-making as well as their degree of influence. The positive coefficients mean that an increase in the feature value increases the predicted outcome and a negative coefficient implies the opposite. This additive and explicit form makes the linear models very useful especially those areas that demand traceable and explainable

ISSN: 3142-788X
PP: 01-12.

decision logic. Another mathematically interpretable formulation is decision tree models which divide the input space into a finite collection of decision regions on the basis of feature based conditions. In this scheme, the prediction is a sum of all terminal regions, with each region having a fixed value of output in the event the input instance is within that region. The use of indicator functions is to identify whether a particular input falls within a particular region and only one region is accepted with any given prediction. This database is also reflective of the logical flow of a tree, with the decision-making process occurring in a sequential order based on a set of simple and easily readable rules. Every line on its way to the leaf represents a clear decision rule, which can be examined and adopted. Formal verification, sensitivity analysis, and regulatory auditing are some of the tasks that do these mathematical representations especially well. Defining the relationship between inputs and outputs explicitly, interpretable formulations enable practitioners to investigate the model behavior given different conditions, identify unwanted biases, as well as, to meet the requirements of safety and fairness. Such models are therefore a sound basis of implementing machine learning systems in high stakes and high risk applications.

3.3. Evaluation Metrics

This is because a balanced score of interpretability and predictive performance is needed to evaluate the interpretable machine learning models, which otherwise transparency threatens operational effectiveness. The metrics based on interpretability quantify the simplicity with which a human being can comprehend and reason concerning the behaviour of a model. An important metric, especially in linear models, is model sparsity which can be defined as the number of active features involved in prediction; more sparse models tend to be easier to interpret because of their reduced cognitive load, not to mention the ability to emphasize the most relevant variables. In decision trees and rule-based systems, rule length and rule depth are usually used metrics, although shorter rules and less deep trees represent simpler and more understandable decision logic. Also, the stability of explanations is becoming widely regarded as a valuable criterion, which measures the consistency of explanations given the similarity of the input or the familiarity with re-training a model. The fact that stable explanations are sensitive to slight variations in input data helps to build trust because they are not sensitivity to critical differences in the interpretation. Traditional performance metrics, as well as interpretability, are necessary as they will authenticate the feasibility of the model. Predictive accuracy is a holistic approach to determine the level of correctness whereas precision and recall are more detailed measures of performance, especially in imbalanced or safety-important datasets when false positives or falses have dissimilar risks. Sensitivity to noise is also tested to determine the sensitivity of the model to perturbation or uncertainty in input data which is extremely important in real-world scenario where data quality can be variant. The evaluation structure is built by examining interpretability and performance measures together, thus so that models in question are not merely transparent and explainable but sound and efficient. This is an elaborate assessment strategy that aids important decision-making in choosing and implementing interpretable models in high-stakes implementations.

WCRG
Global Research
ISSN: 3142-788X
PP: 01-12.

4. RESULTS AND DISCUSSION

4.1. Comparative Performance Analysis

Table 1: Comparative Performance Analysis

Model Type	Accuracy (%)	Interpretability (%)	Suitability for Critical Systems (%)
Linear Models	70%	95%	95%
Decision Trees	85%	85%	85%
Ensemble Models	92%	40%	60%
Deep Neural Networks	95%	20%	40%

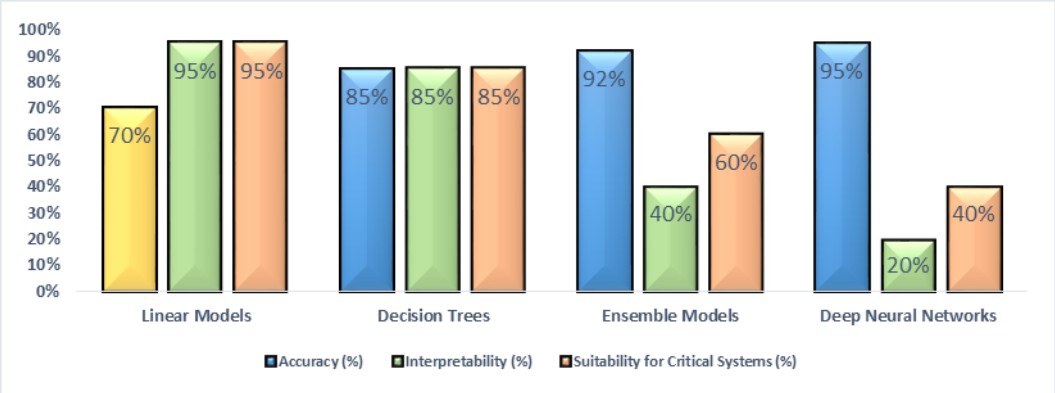

Fig 3 - Comparative Performance Analysis

4.1.1. Linear Models

Linear models prove to have moderate predictive capabilities and provide very high interpretability because decisions are formulated based on weight components of the input features. Their plain mathematical format enables the stakeholders to clearly comprehend, verify and audit model conduct that is essential under controlled and high hazard set-ups. Their transparency and reliability allow them to be very appropriate in critical decision systems where explainability and accountability are the key factors to be found but due to their lack of ability to resolve nonlinear patterns, their accuracy might be less than other systems.

4.1.2. Decision Trees

Decision trees have a good trade off between accuracy and interpretation, and as a result, are often used in applications where this is important. They allow comparatively good predictive performance and a layered form of organisation allows understandability of the decision making logic due to clear rules and conditions. Validation and debugging can be assisted by using the ability to follow a prediction with a series of human comprehensible splits. Consequently, they are ideal in critical systems where performance and transparency is a necessity.

4.1.3. Ensemble Models

Ensemble models are prediction models, including random forests and boosting techniques, which offer high predictive accuracy through the combination process of a few base learners. This improvement in performance is however disadvantaged by the decrease in interpretability as the

ISSN: 3142-788X
PP: 01-12.

combination of several models may blur out the decision-making paths. Although to some extent this problem can be resolved by post-hoc explanation techniques, the complexity of ensembles (in essence) limits their applicability to safety-critical systems. As a result, they tend to be used in those situations, in which the accuracy is both more valued and interpretability needs are moderate.

4.1.4. Deep Neural Networks

Because deep neural networks are able to model complex and high-dimensional nonlinearities, they are highly accurate especially in complex tasks. This advantage is accompanied by their internal representations which are mostly opaque which makes their interpretability very low. Their inability to explain and confirm their decision-making processes restricts their use in the critical systems where transparency, trust and regulation compliance is a binding factor. Therefore, deep neural networks are typically employed in the situations when performance is prioritized over the issues of explainability or when good post-hoc interpretability tools can be used reliably.

4.2. Discussion of Trade-offs

The comparative analysis states that interpretability and predictive performance are not absolute concepts, and it is a continuum that cannot be widely misused depending on the demands of application. The findings show that interpretable models including linear models, and decision trees can reach competitive levels of accuracy with the help of proper feature engineering, domain sensitive variable selection, as well as regularization methods that help control model complexity. These models can learn important patterns in the data avoiding an extremely complicated architecture by emphasizing meaningful properties and removing redundancy. The result contradicts the widely accepted idea that to be a high performer, one must use black-box models and supports the viability of interpretable methods in most real-life contexts. The trade-offs are however magnified with the occurrence of an increase in the complexity of the problems. Simple models are typically outperformed by ensemble methods and deep neural networks when they are applied in problems with nonlinear relationships, with high-level data or with large dimensional features. These types of models provide high levels of accuracy but because they are not highly transparent, there are issues of trust, accountability and regulatory adherence. Such opacity might not be tolerated in safety-critical and regulated areas, even though it can be more efficient. Consequently, the choice of suitable model is required to be informed by a comprehensive evaluation of the risk of application, legal and ethical limiting factors, and end users and decision-makers expectations. In addition, the needs of the stakeholders are very critical in dictating the acceptable compromise between accuracy and interpretability. Automated decisions are usually subject to clear explanations in order to justify and validate them by domain experts, regulators and affected individuals. In this respect, a small decrease in the level of accuracy can be agreed to in favor of a higher degree of transparency and controllability. On balance, the discussion highlights the need to define the model choice based on the context and interpretability must be considered a design requirement, not an ad-hoc effort that allows the implemented systems to prove their effectiveness and quality.

ISSN: 3142-788X
PP: 01-12.

5. CONCLUSION

This paper has conducted an extensive and coherent analysis of interpretable machine learning systems to critical decision systems with focus on their viability in terms of their practical utility and their increasing demand in high stakes areas of application. The paper proves that the concept of transparency may be successfully implemented at every stage of the machine learning lifecycle by studying the principles of interpretability, surveying the methods of providing intrinsic and post-hoc explanations, and suggesting a systematic perspective on the methodology, thus is able to be implemented successfully. As the analysis attests, the concept of interpretability does not have to be regarded as a related or optional feature, but it is essential to the presence of systems that determine safety, fairness, and human wellbeing. These comparative evaluation findings demonstrate that interpretable models, including linear models and decision trees, can produce reliable and operationally viable performance with assistance of cautious feature engineering as well as regularization, together with domain understanding. Although the complex models such as the ensemble models and the deep neural networks can consistently be found to be more accurate, the results suggest that in a number of real-life applications the performance difference can be reduced without necessarily hindering the explainability. The balance is especially critical in the critical decision systems when the accountability, auditability and regulatory compliance matters considerably more than the predictive accuracy. Open system models help stakeholders to see the logic of their decisions, identify biases, and ensure that the systems are operating as intended, creating a sense of trust and supporting responsible implementation. In addition, the proposed framework emphasises the role played by matching the model selection with application risk, regulatory restrictions and the expected results by the stakeholders. Introducing interpretability evaluation, auditory mechanisms and human supervision in the deployment process has made machine learning systems resistant and reliable with time. All of these are necessary to ensure model drift, ethical issues, and informed human-AI collaboration during decision-making. In the future, a number of research lines become critical in developing reliable artificial intelligence. The creation of standardized and common interpretability measures would allow making consistent assessment and comparison of models across domains. Future work can take advantage of the attractiveness of hybrid modeling techniques that merge the ability of complex models to express and the ability of more understandable models to provide insights. Further, more should be done with respect to human-centered assessment issues which will guarantee such explanations are not solely technically right but also significant and usable by the end users. Together, these initiatives will help the responsible deployment of interpretable machine learning systems that will facilitate the safe, fair, and responsible decision-making in critical areas.

REFERENCES

[1] L. Breiman, "Statistical modeling: The two cultures," *Statistical Science*, vol. 16, no. 3, pp. 199–231, 2001.

[2] C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, 2nd ed. Munich, Germany: Leanpub, 2022.

[3] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society: Series B*, vol. 58, no. 1, pp. 267–288, 1996.

ISSN: 3142-788X
PP: 01-12.

[4] J. H. Friedman and B. E. Popescu, "Predictive learning via rule ensembles," *Annals of Applied Statistics*, vol. 2, no. 3, pp. 916–954, 2008.
[5] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.
[6] Y. Lou, R. Caruana, J. Gehrke, and G. Hooker, "Accurate intelligible models with pairwise interactions," in *Proc. 19th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, Chicago, IL, USA, 2013, pp. 623–631.
[7] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. 31st Int. Conf. Neural Information Processing Systems (NeurIPS)*, Long Beach, CA, USA, 2017, pp. 4765–4774.
[8] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 1135–1144.
[9] A. Datta, S. Sen, and Y. Zick, "Algorithmic transparency via quantitative input influence," in *Proc. IEEE Symp. Security and Privacy*, San Jose, CA, USA, 2016, pp. 598–617.
[10] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box," *Harvard Journal of Law & Technology*, vol. 31, no. 2, pp. 841–887, 2018.
[11] Z. C. Lipton, "The mythos of model interpretability," *Queue*, vol. 16, no. 3, pp. 31–57, 2018.
[12] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv:1702.08608, 2017.
[13] R. Guidotti *et al.*, "A survey of methods for explaining black box models," *ACM Computing Surveys*, vol. 51, no. 5, pp. 1–42, 2018.
[14] A. Holzinger, "From machine learning to explainable AI," in *Proc. IEEE World Congress on Services*, Milan, Italy, 2018, pp. 55–66.
[15] E. Samek, T. Wiegand, and K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," *ITU Journal: ICT Discoveries*, vol. 1, no. 1, pp. 1–10, 2017.