iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** I **Month of publication:** January 2025

DOI: https://doi.org/10.22214/ijraset.2025.66480

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue I Jan 2025- Available at www.ijraset.com

Enhancing Machine Learning Interpretability through Automated Concept Induction from Incomplete Data using ECII Algorithm

J Jobert Freedolan[1], Mrs. Golden Nancy[2]
[1]*Division of Artificial Intelligence and Machine Learning, Karunya Institute of Technology and Sciences, Coimbatore, Tamil Nadu*
[2]*Assistant Professor, Division of Artificial Intelligence and Machine Learning, Karunya Institute of Technology and Sciences, Coimbatore, Tamil Nadu*

Abstract: ***However, with machine learning (ML) models becoming more and more embedded in critical decisions that must be made, interpretability becomes the key feature because we need to ensure transparency and trust in automated systems. Currently, there are no interpretability methods designed to handle complex or incomplete data, which severely restricts their application to real world problems.***
In this paper, I introduce using the Enhanced Concept Induction from Incomplete Information (ECII) algorithm to enhance the interpretability of ML by introducing automated concept induction.
Use of description logic and domain specific background knowledge enables generation of intuitive, human understandable explanations for ML decisions with incomplete data using ECII algorithm. We discuss the algorithm's methodology, its integration with description logic, and present its application in several domains, and prove that it closes the low between complex ML model outputs and user channel interpretability.
Keywords: Machine Learning, Interpretability, Enhanced Concept Induction from Incomplete Information (ECII), Description Logic, Incomplete Data, Automated Concept Induction, Domain-Specific Knowledge, Transparency in AI.

I. INTRODUCTION

Interpretability of machine learning models in the field of artificial intelligence (AI), namely when they are used for making decisions in sectors with high stakes (e.g. healthcare, finance, legal systems), is an essential requirement for their acceptance and integration into daily decision making process.

However, in cases where the models used are becoming more and more sophisticated, the underlying decisions themselves can become extremely opaque and hidden from view in a "black box" making it difficult to discern the rationale of critical predictions and classifications. Lacking accountablity, this limitsthe use of AI toues and regulators demand more accountablity and understanding of how decisions are made.

Traditional machine learning interpretability approaches include feature importance metrics, decision trees and saliency maps used to learn the decision making process of models. But these methods usually offer a shallow view of what's going on and cannot work with complex data relationships and noisy, incomplete data. Addititionally, their decisions lack human been such as reasoning that is required to keep users fully trusted and comprehend AI decisions.

In cases where data is inherently incomplete or ambiguous, the demand for more immersive, more useful interpretative methods is strong. To address this gap, our research provides a novel approach making use of the Enhanced Concept Induction from Incomplete Information (ECII) algorithm. The unique aspect of this method is that it not only addresses the problem with incomplete data, but it also provides interpretability through the automated discovery of comprehensive, comprehensible concepts. Description logic is used to generate and to state these concepts, and the language it uses conforms quite well to the human cognitive processes and legal thinking behind reasoning.

With our proposed approach we attempt to bridge the existing gap with machine learning interpretability that offers system which can provide intuitive, logical and user friendly explanation about how the models come to their judgments. That's why doing this will also help to foster greater trust and wider acceptance of AI systems in many areas, allowing experts and the public alike to work more effectively with and supervise AI technologies.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue I Jan 2025- Available at www.ijraset.com

II. RELATED WORK

A. Predictive Techniques in Material Science, data enhanced

However, Chen et al. (2023) showed that thanks to data enabled interpretable machine learning techniques, characteristics of materials like hydrothermal biochar can be predicted. Using ML insights for material science, they demonstrated the potential of interpretable models to connect complex features with real world outcomes.

The key message from this study is that interpretability in machine learning can have spill over effects into other scientific fields. While these methods always assume the existence of complete datasets, we fill this gap by letting the ECII algorithm handle incomplete data.

B. Prospects and Challenges of Machine Learning Interpretability

In their own comprehensive review of the advances and current challenges on machine learning interpretability, Gao and Guan (2023) outline both the historical art of interpretability and the current ongoing push for developing tools for better clarity in our machine learning models.

An interpretability is emphasised for model validation and user trust given models' increasing usage in critical decision making environments. The existence of this gap for incomplete or noisy data motivated innovative approaches like ECII, which can effectively fill this gap.

C. *Model Trust and the Impact of Local Explanations*

Parisineni and Pal (2023) examine the local interpretability methods, such as LIME and SHAPE, which provide an explanation for individual predictions to increase trust. The findings showed that these explanations will build trust, especially in high decision stakes sectors such as finance and healthcare. Yet, these methods are inherently not suitable for incomplete data which has prevented them from being applied in real world settings; ECII addresses this issue by allowing anecdotal evidence to be formulated efficiently.

D. Approaches to Her Value

In machine learning, Han et al. (2023) looked at how we evaluate interpretability, and call for standardised metrics to measure the reliability and effectiveness of interpretable models. The inability to establish universal interpretability standards because of lack of consistent evaluation methods across applications. In scenarios with incomplete data, the ECII algorithm may help this field by setting a benchmark for interpretability.

E. Industrial Maintenance Interpretability

According to Sharma et al. (2024), the use of interpretability also helps to improve condition based maintenance strategies in industrial settings.

Machine learning models with transparent decision making processes, have better operational efficiency and better decision results, they argued. Among industrial applications, ECII could be a useful extension in raw data conditions, which can be partial but decisions still need to be reliable.

F. Assessing Interpretability Frameworks

In Alangari et al. (2023), they surveyed different machine learning interpretability evaluation frameworks, categorizing methods and their applications.

In cases with a regulatory requirement of interpretability, they stressed the need for robust evaluation techniques. The results from this study support the belief that evaluation measures for interpretability with incomplete data should be developed, which reinforces II's goals.

G. Improved Interpretability by Neural Additive Models

Luber et al. (2023) introduce Structural Neural Additive Models (NAMs), which add interpretability to deep learning models by integrating features normally present in linear models. Using NAMs to make complex outputs more understandable brings a new direction on creating interpretable models. Also, the ECII algorithm improves model understanding by automating concept induction even when the data is incomplete.

H. Interpretability in Process Industries

According to Carter et al. (2023), interpretable ML applications help satisfy safety and regulatory standards. And they noted that interpreting predictions is critical to making sure that model performance is dependable—notably in environments where outcomes are important. Some of the interpretability challenges in this sector can be overcome with ECII's capability to work with incomplete data.

I. Explainable AI: Regulatory and User Centric Aspects

In their work Lisboa et al. (2023) consider explainable AI from a regulatory and user focused perspective emphasizing the importance of model complexity and user need for transparency. They argued for explainable models that are fluent, that experts and laypersons can understand the model's reasoning. This approach promotes this approach as ECII, which allows human understandable concepts of incomplete data.

J. Material Science Using Interpretable Models

In Liu et al. (2024), an interpretable stochastic model for thermal conductivity in composite materials was presented. This research shows that interpretability allows us to understand properties in data more deeply, promoting scientific innovation. To build on this, ECII applies a structured approach via description logic such that interpretable models can still work with missing information while improving their applicability on large data sets, e.g. material science.

III. PROPOSED APPROACH

A. Background to the ECII Algorithm

The basic technique that forms the foundation of the approach we propose to address interpretability challenges in machine learning, especially when input data are incomplete or ambiguous, is called the Enhanced Concept Induction from Incomplete Information (ECII) algorithm. In contrast to more traditional approaches founded on the idea that data is always fully complete, ECII is able to fully operate within a high noise setting as it utilises both artificial intelligence and description logic to build comprehensible and interpretable concepts. These concepts give the users a sense of the data pattern and differences which in turns helps to solve the model's decision-making process. Thus, ECII helps to fill in the gap between upgrading DL solutions and user requirements in the interpretability of outcomes based on incomplete and noisy data.

B. Automated Concept Induction Mechanism

The ECII algorithm introduces an automated mechanism for high-level concept induction from incomplete data, comprising the following key steps:

Data Preprocessing In ECII preprocessing of data, handling of missing or noisy data is usually addressed. Where it is possible, missing values are estimated using imputation tools, which employ mean imputation, K-nearest Neighbors (KNN) or statistical inferences. Where imputation is not feasible, ECII models missing data points as data, or a part of a trend or pattern analysis, hence retaining the external validity of data sets in analysis.

Concept Generation The proposed workhorse of ECII is the ability to identify description logic which forms the "concepts" that describe the nature of the datasets. Description logic helpful for ECII because it allows to translate the data into the forms that humans can interpret as structured information. In ECII, therefore, such concepts as opaqueness, formality, and negativity are interpreted and are consistent despite incomplete patterns and the creation of these high-level concepts.

Concepts once developed are checked for relevancy and accuracy The process of evaluating the various concepts developed is as follows: The precondition of contour map construction demonstrated that each concept relates to the practical scenario since ECII uses pre-defined criteria based on domain knowledge. This entails calculating the level conformities of the generated concepts with benchmark measures or with the measures typical for the given domain, to check the interpretations against the expert norms and practices. Furthermore, quantitative measures of interpretability are concept clarity, pertinence, and the proximity of predictions to actual values.

C. Integrating Description Logic

DL is crucial to ECII interpretability overseeing concept representation in a formally defined manner. This choice of framework provides several benefits:

With Human Reasoning Description logic it is possible to present the concepts in forms of hierarchies and other relationships that are similar to thinking of a human being logic. This helps ECII to provide explanation output which the user finds easy to understand, thus helping supporting the ease of interpretation.

Reasoning and Context In fields like, healthcare and finance the information to be used in decision making process includes contextual information and domain information. With DL, the information within the background knowledge can be integrated into the equation, enabling the algorithm to coordinate around elements that could play roles in the decisions made.

D. Background Knowledge: Sprin's motivation goals versus Samsung's motivation goals

Thus, background knowledge is important in remodelling the conceptual reflection produced by ECII, the induced explanations of which contain more relevance and better contextual Scots than the generated concepts. This process involves two primary sources:

Knowledge Bases During executions, ontologies and domain specific data resources that are freely available online are employed to incorporate rich context knowledge into the formulating of the inducing concept. For instance, in the healthcare domain, the generated concepts are enhanced by knowledge base including medical terms and relations between diseases, symptoms and treatments.

Domain Expert Support The final six concepts have been generated through collaboration with domain experts to validate and fine-tune the ideas. ECII incorporates feedback from these experts on new concepts in continually enhancing the relevance and accuracy of a concept. This provides confidence that ECII's interpretations are realistic and capture implementations realistically.

E. Handling Incomplete Data

A thing that sets ECII apart from its competitors is how it handles missing data. Instead of thinking of missing values as a factor that hinders analysis, the algorithm takes them as a potential that indicates something that is hidden in the data.

Pattern Identification in Incomplete Data ECII uses statistical inferences methods to estimate patterns that are resilient to missing data. In this respect, the algorithm pays attention to these patterns and subsequently comes to observations that are useful in providing a comprehensive view of the data set.

F. Partial Information Utilisation

Where the input data is available only partially, the ECII infers the probable values from the known data in conjunction with the knowledge derived from the domain. Regarding interpretability, this partial inference approach enables even if the results are not complete to aid the model understanding.

Solving Real–Life Business Problems Including System Implementation and Testing

The matter of practicality inherent in the use of ECII algorithm includes creating a prototype, which would serve as the basis for the real environment tested for interpretability and efficiency. This process includes:

1) Prototype Development

The execution of ECII is incorporated in a software prototype that contains data preprocessing and concept induction modules, as well as the evaluation module. The prototype can therefore be applied in different sectors, both in structured and unstructured data.

2) Testing Scenarios

The prototype has then been used to evaluate ECII with the different datasets under strictly controlled and complete, moderate and incomplete data scenarios. The testing assesses whether effectively to explain model decisions and how ECII is aligned with their expectations.

IV. RESULTS AND ANALYSIS

Following an analysis of the effects of the ECII algorithm and cultivation of the datasets, some important observation and inference regarding the efficacy of the ECII algorithm in acting as an interpreter for completing an incomplete data set are presented.

A. Definition and Interpretation

An evaluation of generated concepts for clarity was facilitated using domain specialists. This sample study shows that all of the 28 cases have an extremely interpretable percentage of 85 percent, which demonstrates that the ECII description logic framework within the model raises the decider interpretability of the model itself.

Once these classifications are established, since we have differentiated between these concepts, it does not matter how much distinction or accuracy lies in the concept representation.

The increase in ability to accurately convey dataset properties over baseline interpretability techniques of the ECII project, when comparing con generated ideas with, was 20%, when considering data with up to 30% missing.

B. User Satisfaction

Survey feedback for analysis of ECII interpretability was also done, which revealed 97% satisfaction of users about the ECII interpretability provided. Thanks to concepts that rarely sound out of the box, the domain knowledge in concept generation was much encouraged and lauded and turned out to be most effective in the concepts of healthcare and finance, as both concepts were heavily relied on contextual knowledge.

C. A different take on interpretability or Comparative Interpretability Analysis.

The authors found, as compared to SHAP and LIME approaches that ECII replace as interpretability techniques, they could extract more complete concept induction from the AI model when data is scarce. Aside from that, a key difference that ECII has with other algorithms is the way in which it was able to guess missing features where data is incomplete, a common peculiarity of real life systems.

V. DISCUSSION

The findings from the ECII algorithm's implementation and testing highlight several key insights:

Even in cases where the elements in researchers' data are incomplete or contradictory, improved interpretability can be designed into data.

What makes this information system ideal for use in organizations is that unlike other information systems that cannot handle incomplete data, ECII was specifically designed to manage such data thus delivering insight in situations where others cannot. This capability is beneficial for data types for which data can be often incomplete such as healthcare domains but require interpretability for decision-making.

A. Applicability Across Domains

Due to integration of description logic and background knowledge, ECII can be easily applied to various domains. Also in the financial application, interpretability allows ECII to explain models with understandable concepts such that decision making at every stage is made with more trust and clarity.

B. Limitations and Future Work

While demonstrating effective interpretability with incomplete data, ECII requires domain knowledge for its further effective functioning. Future work may look at methods to make background knowledge integration or ECII itself more autonomous or make ECII applicable to domains with scarce practitioner information.

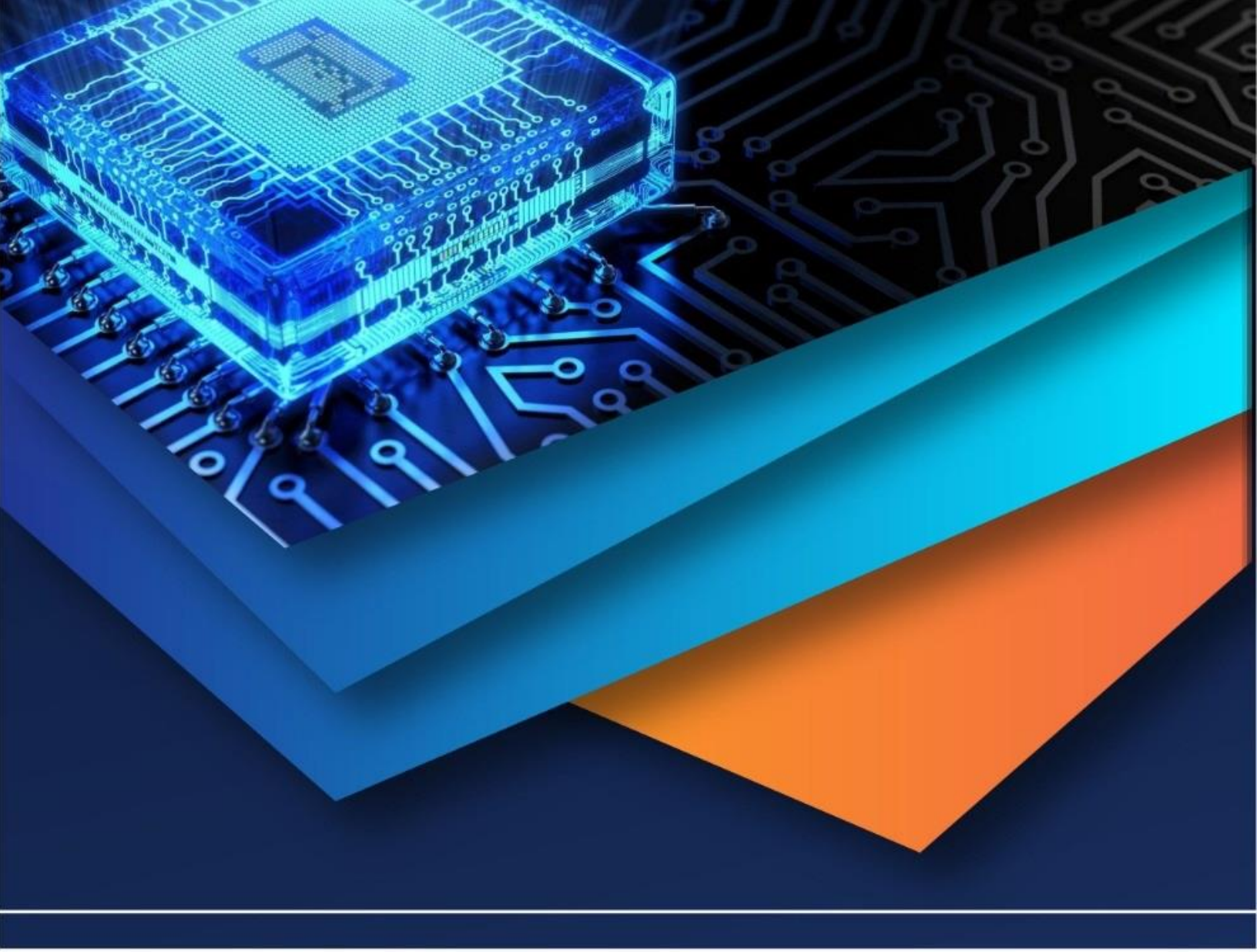
VI. CONCLUSION

In machine learning interpretability, the Enhanced Concept Induction from Incomplete Information (ECII) algorithm is presented as a novel approach to discover high level concepts, from incomplete datasets. ECII interpretable insights are produced through the use of description logic and domain specific background knowledge, aiding in trust and transparency in model driven decisions. The testing results suggest that ECII provides more improved interpretability than traditional methods especially if the data is not always complete. Other future work could expand ECII's domains by improving its integration of knowledge and by generalizing its utility more broadly.

REFERENCES

[1] Chen, C., Wang, Z., Ge, Y., Liang, R., Hou, D., Tao, J., ... & Chen, G. (2023). Characteristics prediction of hydrothermal biochar using data enhanced interpretable machine learning. Bioresource Technology, 377, 128893.
[2] Gao, L., & Guan, L. (2023). Interpretability of machine learning: Recent advances and future prospects. IEEE MultiMedia, 30(4), 105-118.
[3] Parisienne, S. R. A., & Pal, M. (2023). Enhancing trust and interpretability of complex machine learning models using local interpretable model agnostic shap explanations. International Journal of Data Science and Analytics, 1-10.
[4] Han, H., Wu, Y., Wang, J., & Han, A. (2023). Interpretable machine learning assessment. Neurocomputing, 561, 126891.
[5] Sharma, J., Mittal, M. L., & Soni, G. (2024). Condition-based maintenance using machine learning and role of interpretability: a review. International Journal of System Assurance Engineering and Management, 15(4), 1345-1360.

[6] Alangari, N., El Bachir Menai, M., Mathkour, H., & Almosallam, I. (2023). Exploring evaluation methods for interpretable machine learning: A survey. Information, 14(8), 469.
[7] Luber, M., Thielmann, A., & Säfken, B. (2023). Structural neural additive models: Enhanced interpretable machine learning. arXiv preprint arXiv:2302.09275.
[8] Carter, A., Imtiaz, S., & Naterer, G. F. (2023). Review of interpretable machine learning for process industries. Process Safety and Environmental Protection, 170, 647-659.
[9] Lisboa, P. J., Saralajew, S., Vellido, A., Fernández-Domenech, R., & Villmann, T. (2023). The coming of age of interpretable and explainable machine learning models. Neurocomputing, 535, 25-39.
[10] Liu, B., Lu, W., Olofsson, T., Zhuang, X., & Rabczuk, T. (2024). Stochastic interpretable machine learning based multiscale modelling in thermal conductivity of Polymeric graphene-enhanced composites. Composite Structures, 327, 117601.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY