Print ISSN: 0976-2876 Online ISSN: 2250-0138

Available online at: http://www.ijsr.in

INDIAN JOURNAL OF SCIENTIFIC RESEARCH

DOI:10.32606/IJSR.V15.I1.00012

Received: 18-05-2024 Accepted: 21-07-2024 Publication: 31-08-2024

Indian J.Sci.Res. 15 (1): 113-118, 2024 Original Research Article

ENHANCING DRUG DISCOVERY EFFICIENCY THROUGH EXPLAINABLE AI: TRANSITIONING FROM BLACK-BOX TO WHITE-BOX MODELS

SHALINI LAMBA[a1], PRATHAM GUPTA[b] AND SHREY NAGAR[c]

[abc]Department of Computer Science, National P.G College, Lucknow, Uttar Pradesh, India

ABSTRACT

Technology has continued to transform the educational landscape, also the use of A.I. in the healthcare sector has gained prominence. We all know how complex our healthcare systems are. AI (Artificial Intelligence) can transform these fields and has already started to transform this field. AI's inclusion will make it very easy for doctors to monitor patients, and it will be straightforward to detect diseases they are suffering from. If AI tools are integrated into this field, we will need healthcare providers with essential knowledge and tools. This research paper will explore the diverse applications of AI in this field, ranging from imaging analysis and predictive analysis. We will be discussing how AI can positively affect the medical field and what methods it can use to increase the performance of AI in this field. In this paper, mainly, we will be discussing the idea of using Explainable Artificial Intelligence (XAI) in this field to improve performance and gain maximum efficiency by making drug discovery model from black box model to white box model. We will be discussing all these topics which are being used or can be used in this sector.

KEYWORDS: Artificial Intelligence, Healthcare, Explainable AI (XAI), Machine Learning, White Box Model (WBM)

Artificial intelligence works in many ways, using principles and tools such as mathematics, logic, and biology. One of the important features of today's AI devices is their ability to understand a wide variety of information, such as text and images. Machine learning has been the most popular form of AI in recent years and forms the basis of many applications today. Instead of following predefined instructions, machine learning allows machines to discover patterns and create their own rules when taking in new information and knowledge.

In healthcare, AI has advanced medicine in ways that were previously unimaginable, opening up new possibilities by solving some of the world's health challenges. For example, the cutting-edge AI-powered protein structure prediction algorithm AlphaFold solves the protein folding problem that has held back major advances in biology and medicine for the past 50 years.

AI is being used or tested for a variety of medical and research purposes, including disease diagnosis, chronic disease management, healthcare delivery, and drug discovery (Figure 1). AI has the potential to help solve important health problems, but may be limited by the quality of health data and the lack of AI's ability to inform some human characteristics.

AI systems work using algorithms and data. First, large amounts of data are collected and used for mathematical models, or algorithms, that use the data to identify patterns and make predictions, in a process called training. Once trained, algorithms are used in multiple applications, constantly learning and adapting to new data. This allows AI systems to perform complex tasks like image recognition, language processing, and data analysis with greater accuracy and efficiency over time (Alowais *et al.* 2023; Hassija *et al.* 2023; Nuffield Council on Bioethics, 2018; Khalid *et al.* 2023).

What is Explainable AI (XAI)

Explainable Artificial Intelligence (XAI) is a branch of AI that focuses on making complex AI models understandable and customizable for effective decision-making, particularly in science and technology. Currently, AI is widely used in the medical and healthcare fields. This includes predicting future illnesses from genomic data, diagnosing critical diseases early by analysing electronic health records, recognizing serious conditions in advance through AI-based Early Warning Scores, and using smart clinical decision-support models.

The challenges related to adaptability, transparency, and bias in current AI models have led to a push for more transparency in AI design. The goal is to help professionals better understand and customize existing AI mechanisms through explainable AI models, ensuring transparency using various techniques. Current healthcare models face limitations when dealing with real-world complexities and uncertainties. Medical decision-support systems need to be robust and precise, as they directly impact human survival. Therefore, AI

[1]**Corresponding author**

models in healthcare must be reliable enough to make appropriate decisions.

There is growing interest in XAI approaches that can explain AI models to solve explainability challenges in various applications. XAI aims to simplify machine learning models by providing clear information about why a model took a particular action, while also developing models that are understandable and transparent without compromising on performance.

The need for more transparent and interpretable AI models, especially in fields like healthcare and medicine, has motivated current research. Many powerful approaches, often referred to as "black boxes," lack transparency, which can be problematic, especially in medical applications where the effectiveness of a diagnosis or treatment is closely linked to the approach used. This research aims to discuss various explainable approaches, tools, and case studies to help young researchers understand the models used in healthcare.

The main contributions of our research include using XAI in drug discovery so that we can convert the process of creating drug from black box model to white box model, helping in tracing the properties of the drug so that we can get the required drug.

Expanding the model's capabilities has various implications, even though AI algorithms have been shown to outperform humans in some tasks. To gain the confidence of healthcare professionals and other stakeholders, predictive models and smart diagnosis approaches must be transparent, understandable, and explainable (Figure 2). The design of these models must clearly explain the decision-making process and the reasons behind specific actions. These frameworks, often referred to as "white-box models**,**" provide a clear sequence of actions (Alizadehsani *et al.* 2024; Srinivasu *et al.* 2022; Nguyen *et al.* 2021).

Shap Model

SHapley Additive exPlanations (SHAP) is a strategy utilized to clarify the yield of machine learning models by deciding how much each input include contributes to a specific expectation. SHAP leverages Shapley Values, a concept from diversion hypothesis, to evaluate the commitment of each feature.

LIME

LIME (Locally Interpretable Model-agnostics Explanations) is a tool that explains how a machine learning model makes a particular prediction for a situation. It helps to understand which features (or components) have the greatest impact on the prediction, especially in the distribution of activities. It also shows the result for a participant in all classes. LIME presents these interactions in an easy-to-understand visual format. This means that LIME may miss some important details about how the features in the original model interact, as it only focuses on the simplified version. Also, the interpretation of LIME can vary depending on the model used, meaning that the same data can lead to different results if different models are used.

Black Box Model

A Black Box Model in XAI refers to machine learning models that operate invisibly, where the inner workings of the model are not easily accessible or interpretable. These models make predictions based on input data, but the decision-making process and the logic behind the predictions are not transparent to users. This lack of transparency makes it difficult for users to understand the behaviour of the model, identify biases or errors, or hold the model accountable for its decisions as shown in Figure 3.

Algorithm: Converting Black-Box Model to White-Box Model for Drug Discovery using XAI (Figure 4)

Input

- Black-Box Model (BBM): A pre-trained AI model used in drug discovery (e.g., predicting drug efficacy or toxicity).
- Drug Data (DD): Data including chemical structure, biological activity, genomic data, etc.
- Explainable AI Tools (XAI Tools): Tools such as SHAP, LIME, and other interpretability methods.

Output

- White-Box Model (WBM): A more interpretable AI model with explanations for decisions.
- Model Interpretations (MI): Insights into how the model makes decisions.

Steps

1. Data Pre-processing (Pre-process Data):
- Standardize and normalize the drug data (DD).
- Identify and encode relevant features (e.g., molecular descriptors, bioactivity data).
- Split the data into training, validation, and test sets.
2. Model Evaluation (Evaluate BBM):
- Evaluate the performance of the black-box model (BBM) using standard metrics (e.g., accuracy, precision, recall, F1-score).
- Store baseline performance metrics for comparison.

3. Feature Importance Analysis (Analyse Features):

- Apply XAI tools (e.g., SHAP, LIME) to the black-box model to determine feature importance.
- Rank features based on their influence on model predictions.

4. Model Decomposition (Decompose BBM):

- Decompose the black-box model by isolating key components:
- Local Interpretations: Use LIME to understand model behaviour for individual predictions.
- Global Interpretations: Use SHAP to understand overall feature impact on the model.

5. Generate Explanations (Generate Explanations):

- For each prediction, generate explanations using the selected XAI tools.
- Store these explanations to compare with model outputs.

6. White-Box Model Design (Design WBM):

- Use the insights gained from the feature importance analysis and model decomposition to design a white-box model.
- Choose interpretable model architectures (e.g., decision trees, rule-based systems, linear models).
- Incorporate the most important features identified by XAI tools into the model.

7. Model Training (Train WBM):

- Train the white-box model using the training dataset.
- Ensure the model remains interpretable while achieving comparable performance to the black-box model.

8. Model Explanation Integration (Integrate Explanations):

- Integrate the explanations generated in step 5 into the white-box model.
- Ensure the model outputs not only predictions but also explanations for each prediction.

9. Model Validation (Validate WBM):

- Evaluate the white-box model on the validation dataset.
- Compare the performance metrics with those of the black-box model to ensure the white-box model is competitive.

10. Refinement and Tuning (Refine and Tune):

- Refine the model by tuning hyper parameters, adding/removing features, or adjusting model complexity.
- Ensure that interpretability and performance are balanced.

11. Final Testing (Test WBM):

- Test the final white-box model on the test dataset.
- Evaluate the interpretability of the model and the quality of explanations provided.

12. Output Model and Explanations (Output Model):

- Output the final white-box model along with the generated interpretations and explanations.
- Document the insights and decision-making processes embedded in the model.

MATERIALS AND METHODS

The methodology for integrating XAI into healthcare AI models involves several key steps:

Data Collection and Pre-processing: Data is gathered from various sources, standardized, and split into training, validation, and test sets.

Model Development: A black-box AI model is developed using advanced algorithms for tasks such as predicting drug efficacy or diagnosing diseases.

Feature Importance Analysis: Tools like SHAP and LIME are used to analyze the importance of different features in the model's predictions.

Model Decomposition: The black-box model is decomposed to generate both local and global interpretations of its decision-making process.

White-Box Model Design: Insights from the previous steps guide the design of a more interpretable white-box model.

Model Validation: The white-box model's performance is validated against the black-box model using standard metrics.

Integration of Explanations: The final model outputs predictions along with explanations, enhancing trust among healthcare professionals.

RESULTS

After applying the XAI tool to the black box model, we identified the key features that affect the model prediction. By combining this information, we created a white box that not only maintains the original black box model’s specificity but also provides more definition. The white box model provides a clear understanding of the specific features that affect the efficacy and safety of drug candidates. The predictive model is based on the black box model but has the advantage of being easily interpretable by researchers and practitioners. This description is important to ensure that the drug discovery process is not only accurate but also reliable and trustworthy (Figure 4).

DISCUSSION

The transition from a black-box to a white-box approach to drug discovery presents several significant challenges to the field, including the need for transparency and the ability to trace causality behind model prediction. Extending white-box models enables more informed decisions and reduces the risk of error and bias that can result from using opaque models. A powerful strategy for overall performance. Free-box models provide clear explanations for predictive models, facilitating better collaboration between AI systems and human experts, leading to better outcomes and treatment outcomes. They increase transparency and accountability while meeting the performance standards of traditional intelligence standards. This advancement represents a significant step in making AI-enabled drug discovery more reliable, easier to understand, and ultimately more effective (Saeed and Omlin, 2023; Hulsen 2023; Salih *et al.* 2024; Love *et al.* 2023).

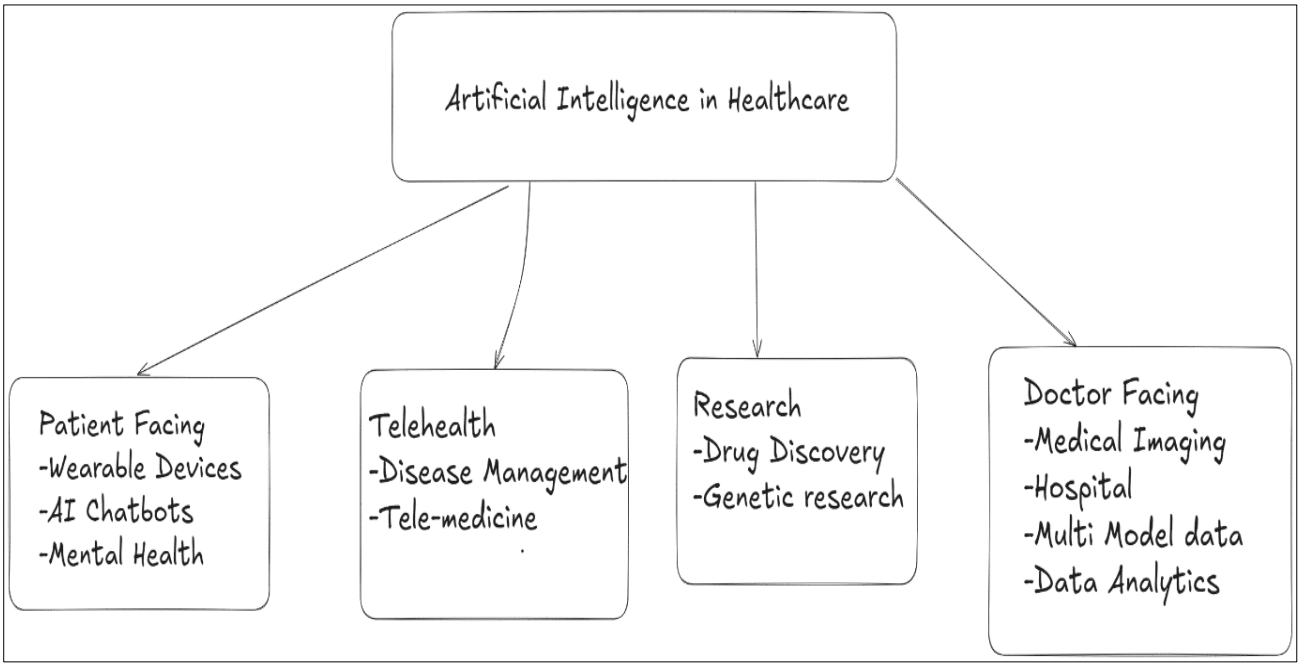

Figure 1: Applications of AI in different fields of healthcare

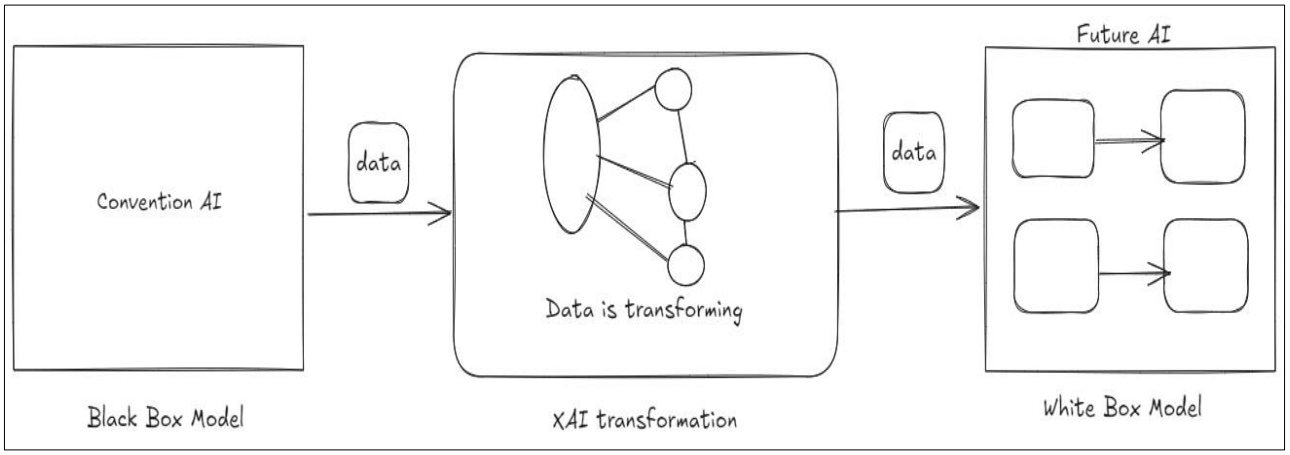

Figure 2: Concept used for converting black box drug discovery model to white box

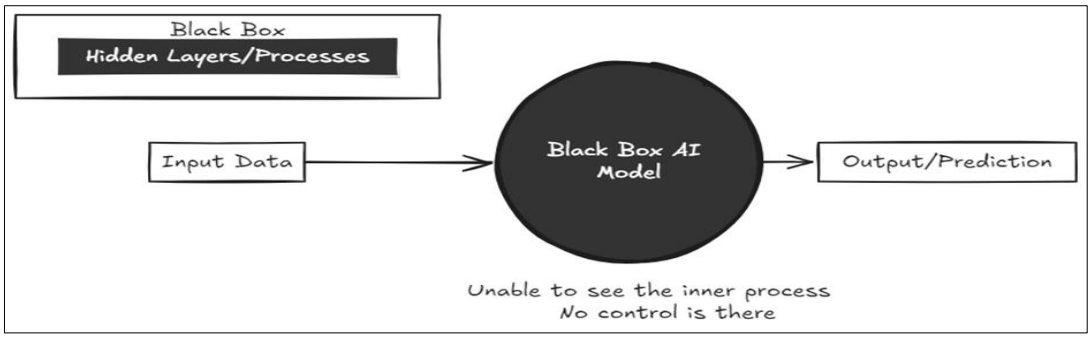

Figure 3: Working of black box model

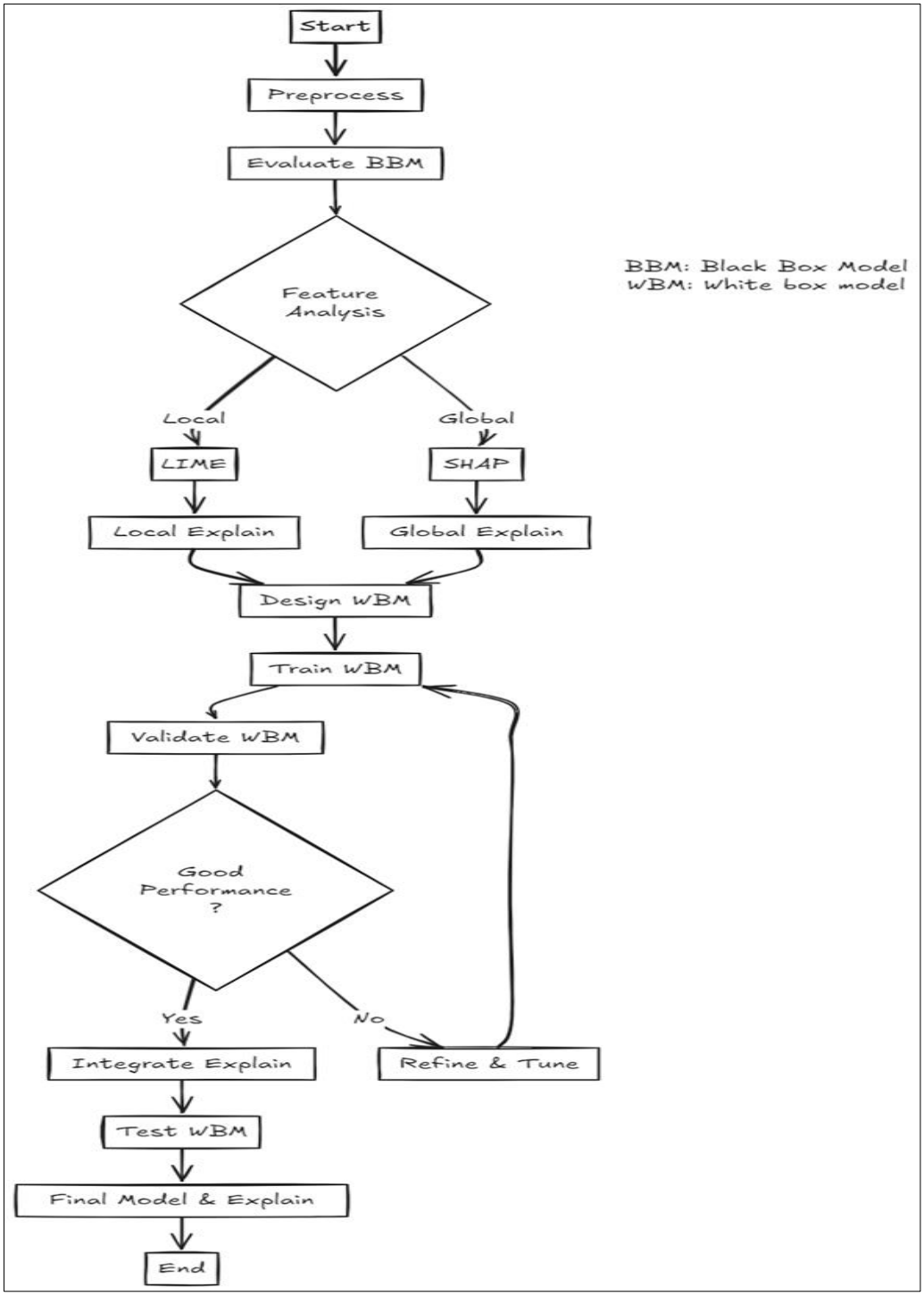

Figure 4: Flowchart for algorithm

ACKNOWLEDGEMENT

I would like to express my heartfelt gratitude to Dr. Shalini Lamba, Head of the Department of Computer Science at National P.G. College, Lucknow, for her invaluable guidance and unwavering support throughout this research. Her insightful suggestions and encouragement have been instrumental in the successful completion of this work. I also extend my sincere thanks to the Department of Computer Science and National P.G. College for providing the necessary resources and a conducive environment for my research. Their continuous support and encouragement have been greatly appreciated.

REFERENCES

Alowais S.A., Alghamdi S.S., Alsuhebany N., Alqahtani T., Alshaya A.I., Almohareb S.N., Aldairem A., Alrashed M., Saleh K.B., Badreldin H.A. and Al Yami M.S., 2023. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC, **23**: 689.

Alizadehsani R., Oyelereb S.S., Hussain S., Calixto R.R., Albuquerque V.H.C., Roshanzamir M., Rahouti M. and Jagatheesaperumal S.K., 2024. Explainable Artificial Intelligence for Drug Discovery and Development - A Comprehensive Survey. IEEE, **12**: 35796-35812.

Hassija V., Chamola V., Mahapatra A., Singal A., Goel D., Huang K., Scardapane S., Spinelli I., Mahmud M. and Hussain A., 2023. Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. Springer.

Hulsen T., 2023. Explainable Artificial Intelligence (XAI): Concepts and Challenges in Healthcare. Department of Hospital Services & Informatics, Philips Research, 5656 AE Eindhoven. **4**(3): 652-666. https://doi.org/10.3390/ai4030034

Khalid N., Qayyum A., Bilal M., Al-Fuqaha A. and Qadir J., 2023. Privacy-preserving artificial intelligence in healthcare: Techniques and applications. Computers in Biology and Medicine, **158**: 106848.

Love P.E.D., Fang W., Matthews J., Porter S., Luo H. and Ding L., 2023. Explainable artificial intelligence (XAI): Precepts, models, and opportunities for research in construction. Science Direct, **57**. https://doi.org/10.1016/j.aei.2023.102024

Nguyen H.T.T., Nguyen K.V.T., Pham N.D.K. and Cao H.Q., 2021. Evaluation of Explainable Artificial Intelligence: SHAP, LIME, and CAM. FPT AI Conference.

Nuffield Council on Bioethics, 2018. Artificial intelligence (AI) in healthcare and research. 28 Bedford Square, London WC1B 3JS.

Saeed W. and Omlin C., 2023. Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. Knowledge-Based Systems, **263**: 110273.

Salih A.M., Raisi-Estabragh Z., Galazzo I.B., Radeva P., Petersen S.E. and Lekadir K., 2024. A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME. https://doi.org/10.1002/aisy.202400304

Srinivasu P.N., Sandhya N., Jhaveri R.H. and Raut R., 2022. Blackbox to Explainable AI in Healthcare: Existing Tools and Case Studies. https://doi.org/10.1155/2022/8167821