IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 27 February 2023, accepted 16 April 2023, date of publication 1 May 2023, date of current version 31 May 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3271635

TOPICAL REVIEW

Explainable Artificial Intelligence for Patient Safety: A Review of Application in Pharmacovigilance

SEUNGHEE LEE[1], SEONYOUNG KIM[1], JIEUN LEE[1], JONG-YEUP KIM[1,2], MI-HWA SONG[3], AND SUEHYUN LEE[4]

[1]Healthcare Data Science Center, Konyang University Hospital, Daejeon 35365, Republic of Korea
[2]College of Medicine, Konyang University, Daejeon 35365, Republic of Korea
[3]School of Information and Communication Sciences, Semyung University, Jecheon 27136, Republic of Korea
[4]College of IT Convergence, Gachon University, Seongnam 13120, Republic of Korea

Corresponding authors: Mi-Hwa Song (mhsong@semyung.ac.kr) and Suehyun Lee (leesh@gachon.ac.kr)

This work was supported by the Korea Health Technology Research and Development Project through the Korea Health Industry Development Institute (KHIDI) funded by the Ministry of Health and Welfare, Republic of Korea, under Grant H I21C1074.

ABSTRACT Explainable AI (XAI) is a methodology that complements the black box of artificial intelligence, and its necessity has recently been highlighted in various fields. The purpose of this research is to identify studies in the field of pharmacovigilance using XAI. Though there have been many previous attempts to select papers, with a total of 781 papers being confirmed, only 25 of them manually met the selection criteria. This study presents an intuitive review of the potential of XAI technologies in the field of pharmacovigilance. In the included studies, clinical data, registry data, and knowledge data were used to investigate drug treatment, side effects, and interaction studies based on tree models, neural network models, and graph models. Finally, key challenges for several research issues for the use of XAI in pharmacovigilance were identified. Although artificial intelligence (AI) is actively used in drug surveillance and patient safety, gathering adverse drug reaction information, extracting drug-drug interactions, and predicting effects, XAI is not normally utilized. Therefore, the potential challenges involved in its use alongside future prospects should be continuously discussed.

INDEX TERMS Machine learning, pharmacovigilance, explainable artificial intelligence.

I. INTRODUCTION

The World Health Organization defines pharmacovigilance (PV) as the science and activities related to the detection, assessment, understanding, and prevention of adverse effects or other drug-related problems [1].

Recent artificial intelligence-based technologies can be an efficient complement to traditional PV methods, which can be costly and time-consuming and can result in adverse drug reactions (ADRs) that go unreported to healthcare professionals.

Artificial intelligence (AI) can improve PV, but its use in PV is still in the early stages of research. Various machine learning (ML) techniques, together with natural language processing and data mining, can be applied to electronic health records, claims databases and social media data to improve the characterization of known drug side effects and reactions, and to detect new signals [2], [3].

AI-based technologies have been criticized for their inexplicable algorithms, despite their high predictive power. In critical decision areas such as healthcare, the reasoning behind a decision is as important as the decision itself, which is why there is growing interest in and research and development around Explainable Artificial Intelligence (XAI).

XAI was developed to improve the transparency of AI systems and generate explanations for them, and seeks to increase trust and understanding by assessing the strengths and limitations of existing models [4], [5], [6]. Approaches that extract information from a model's decision-making process, such as post-hoc explanations, can provide useful

The associate editor coordinating the review of this manuscript and approving it for publication was Jerry Chun-Wei Lin.

information for practitioners and users interested in case-by-case explanations rather than the internal workings of a model [7].

XAI increases the explainability and transparency of AI algorithms by making it possible to interpret the variables that influence decisions, complex internal features, and learned decision paths within a decision process [8], [9]. I.R. Ward et al. successfully quantified the importance of features using an XAI algorithm, further demonstrating the potential contribution of XAI to PV monitoring [10].

The importance of PV in medicine is relevant to all species affected by medical interventions, and ensuring medical safety requires attention and research into approaches such as drug safety reporting and the exchange of reliable and timely information on PV activities [11]. The global pharmacovigilance and drug safety software market size was valued at USD 6.9 billion in 2021 and is estimated to expand at a compound annual growth rate (CAGR) of 10.5% between 2022 and 2030 (Source: www.grandviewresearch.com).

The aim of this study was to review the literature on the use of XAI in PV by identifying publications related to ML/AI and drugs and the rationale for the reported findings. From the perspective of AI and XAI usage, these studies were analyzed, and the findings were summarized, in which the use of XAI in the field of PV is referred to as "PV XAI". The main contributions are highlighted and discussed below:

- This study is clearly an early attempt to review XAI research in PV. Unlike other fields, we found that XAI research in PV is at an early stage of development, limited to a few articles and some methodologies.
- Nevertheless, we have identified the positive potential of PV XAI for drug therapy, ADRs, polypharmacy and drug repurposing.
- While safety issues in real-world healthcare settings may limit the growth of the field, we expect PV XAI research to expand as it has in other areas, and we encourage collaboration and ongoing research discussions with experts in the field.

II. LITERATURE SURVEY

In this study, the trend of XAI in the field of PV was examined. However, the trend was also explored broadly to more diverse aspects, including interpretable artificial intelligence. Although there is a clear difference between Explainable AI (knowledge about what different nodes represent and their importance to model performance) and Interpretable AI (ability to determine cause and effect in a machine learning model), based on the same aim, they were comprehensively reviewed.

There has been a surge in XAI studies in drug-related applications since 2019, with relatively few studies from 2013 to 2018 (Fig. 1). The limited number of publications indicates a demand for more research on XAI in PV applications.

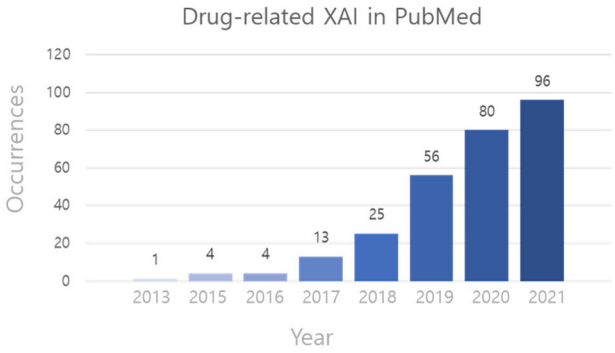


FIGURE 1. Growth of PV articles from 2013 to 2021. "Drug" [All Fields] AND ["eXplainable" [All Fields] OR "Interpretable" [All Fields] OR "Transparent" [All Fields] AND "Artificial intelligence" [MeSH] OR "Machine learning" [MeSH] OR "Deep learning" [MeSH] OR "XAI.

The selection of appropriate search terms for the exploration of XAI-related research in PV was not easy; we started manually with broad keywords. The following five searches were performed: pharmacovigilance XAI (47), pharmacovigilance "explainable artificial intelligence" (76), pharmacovigilance explainable AI (230), pharmacovigilance explainable ML (181), and pharmacovigilance explainable machine learning (213). These search terms were used in a Google Scholar search on 22 June 2022, and the numbers in parentheses are the number of articles returned from each search. Retrieved articles were first screened for titles and abstracts to exclude duplicates, then articles were added through a first full-text review for relevance and a second full-text review based on a selective methodology, resulting in a final selection of 25 unique publications (Fig. 2, Supplementary Table 1).

III. RESULTS

This highlight focuses on PV and reviews of recently developed machine-learning models to predict the explanatory potential and effectiveness of XAI (Fig. 3).

A. ASPECTS OF ARTIFICIAL INTELLIGENCE USE IN PHARMACOVIGILANCE

1) DATASET

a: CLINICAL DATA

PV involves the collection and analysis of data from patient records or other sources to identify causal relationships between medicines and adverse drug reactions. While the potential for monitoring and preventing adverse drug reactions through clinical data is great, exploiting it can be time-consuming because clinical data is highly heterogeneous in structure and domain, and is often managed in multiple files. In addition, specific expertise is required to make sense of clinical information and often requires an understanding of appropriate related prescriptions, which in most cases requires a long preparation time with clinical researchers [36].

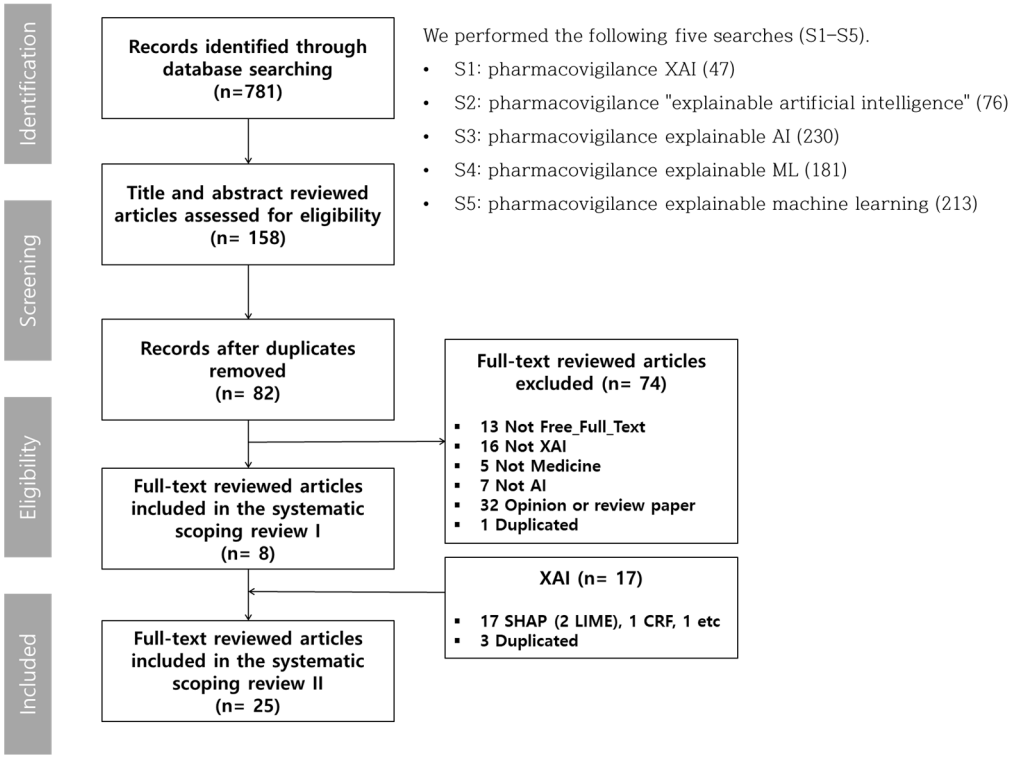


FIGURE 2. The PRISMA diagram depicts the number of records identified, included and excluded, and the reasons for exclusions in PV-XAI.

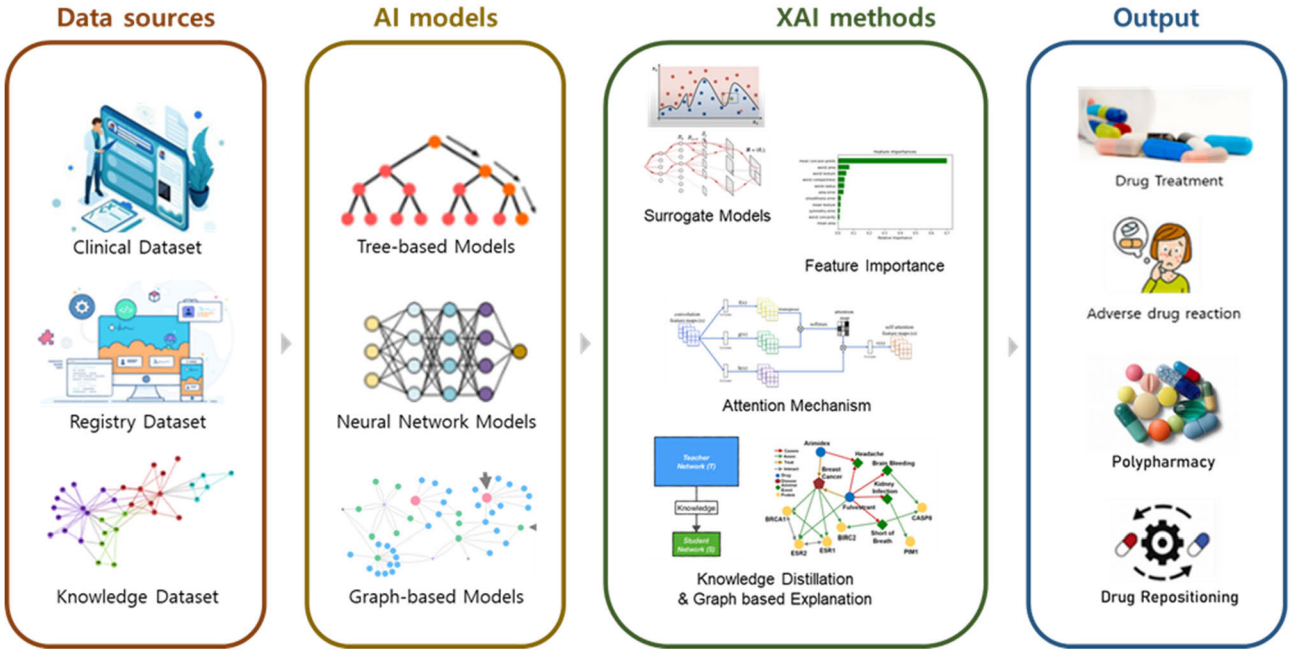


FIGURE 3. Graphical summary of recent studies on pharmacovigilance using XAI.

b: REGISTRY DATA

Data registries are used to evaluate and improve outcomes for populations defined by a condition, disease, or specific type of exposure. Data registries use observational research methods to collect and harmonize data on the treatment, outcomes, and well-being of patients receiving care over time. They also allow large data sets to be aggregated and analyzed for trends or patterns in treatment and outcomes [37].

TABLE 1. Summary of the characteristics of publications included in the analysis.

APPLICATION	DATA SET	CLINICAL CONDITION	ML/DL Models	XAI Category	Model-specific/ Model-agnostic	XAI Method	Ref
ADR	NER datasets	5 drugs	Deep Neural Network	Attention mechanism	Model-specific	Muti-task learning, NER with Attention	[12]
	Swedish Health Record Research Bank, Health Bank at Stockholm university	Structure-Activity Relationships	Deep Neural Network	Attention Mechanism Surrogate Model	Model-specific Model-agnostic	RNN, RETAIN SHAP	[13]
	PharmGKB, DrugBank, SIDER, Gene Atlas, UniProt, KEGG	DILI	Tree-based Graph based	Graph based Explanations	Model-specific	Interpretable Knowledge graph by experts	[14]
	administrative data	ATC_M, ATC_C	Tree-based	Feature Importance Surrogate Model	Model-specific Model-agnostic	MDI, MDA, LIME, SHAP	[10]
	Twitter, PubMed	Opioid Overdose	Deep Neural Network	Surrogate Model	Model-agnostic	LIME	[15]
	SIDER4.1, PubChem	Chemical structure	Graph Neural Network	Surrogate Model	Model-agnostic	SHAP	[16]
	SIDER	Single-meal scenario	Tree-based	Surrogate Model	Model-agnostic	SHAP	[17]
	Children's Hospital of CMU	Antibacterials	Tree-based	Surrogate Model	Model-agnostic	SHAP	[18]
Drug repurposing	21 public databases	Alzheimer's Disease.	Graph Neural Network	Knowledge Distillation & Graph based Explanations	Model-specific Model-agnostic	GraphMask Path based explanations Domain Explanation	[19]
	NCI-DTP data repository	Anticancer drug	Deep Neural Network	Surrogate Model	Model-specific	LRP	[20]
Drug treatment	ChEMBL	BC cell line compounds	Deep Neural Network Graph Neural Network	Attention mechanism Surrogate Model	Model-agnostic	SHAP	[21]
	Ontology	Antibiotics	Deep Neural Network	Graph based Explanations	Model-specific	Ontology preference visualization	[22]
	Individual Case Safety Reports, Signals of Disproportionate Reporting	NSAIDs	Tree-based	Feature Importance Surrogate Model	Model-agnostic	SHAP	[23]
	Cohort study(Finland)	Alzheimer(N06A, N05A, C08, C10 …)	Tree-based	Surrogate Model	Model-agnostic	SHAP	[24]
	Cerner's Health Facts database	Opioid	Graph Neural Network	Surrogate Model	Model-agnostic	SHAP	[25]
	Protein Data Bank	Protein	Graph Neural Network	Surrogate Model	Model-agnostic	SHAP	[26]
	Hospital of soochow university, IWPC open data	Breast Cancer	Tree-based	Feature Importance Surrogate Model	Model-agnostic	SHAP	[27]
	Kaggle Drug classification dataset	5 drugs	Tree-based	Surrogate Model	Model-agnostic	SHAP	[28]

TABLE 1. *(Continued.)* **Summary of the characteristics of publications included in the analysis.**

	Optum	NASAIDs	Tree-based	Surrogate Model	Model-agnostic	SHAP	[29]
	therapeutics-centered network	Insulin	Tree-based	Surrogate Model	Model-agnostic	SHAP	[30]
	STITCH, hEMBL	339 compounds	Tree-based	Feature Importance Surrogate Model	Model-agnostic	SHAP	[31]
	DrugBank, Zerifin TIS	Teratogenic drugs	Tree-based	Surrogate Model	Model-agnostic	SHAP	[32]
	V.A. ANMRC	Sibutramine, Liraglutide, Metformin	Tree-based	Surrogate Model	Model-agnostic	SHAP	[33]
	PRAEGNANT	Medication therapy	Deep Neural Network	Surrogate Model	Model-specific	LRP	[34]
Polypharmacy	DrugBank, ChEMBL	38 drugs, 39 cancer cell line	Deep Neural Network	Attention mechanism	Model-specific	TranSynergy SA-GSEA	[35]

c: KNOWLEDGE DATA

With the development of graph mining, knowledge-based data is emerging as a source of drug research, especially in the pre-marketing phase. Data on drug chemistry [38], drug targets [39], drug side effects [40], biological pathways [41], protein interactions [42], and drug interactions [43] are the most popular knowledge-based datasets used in PV research. H. Kwak et al. reported that integrating the graph structure of knowledge bases with real-world data (RWD) improves the causal interpretability of adverse drug event (ADE) detection [44]. These databases are also accessible through open platforms to be utilized in pharmacovigilance research (Supplementary Table 2).

2) MODELS

a: TREE-BASED MODELS

Tree-based algorithms are conceptually simple but powerful ML methods that are effective on small and large datasets to solve linear and nonlinear modelling problems. Several tree-based ML algorithms were used in this study, including regression trees and ensemble learning, such as random forest, extreme gradient boosting (XGBoost) and adaptive boosting (AdaBoost) [45].

Tree-based ML models with XAI techniques can act as an 'early warning system' for per-patient disease-related adverse outcomes and harmful drugs in a PV monitoring system. With appropriate infrastructure and additional clinical data, these algorithms could provide an autonomous method for monitoring adverse outcomes from medications at the population level, providing a valuable addition to the existing statistical techniques, such as mean decrease of impurity (MDI) and mean decrease in accuracy (MDA), which are currently used and would be the next step in progressing towards a real-time PV monitoring system.

b: NEURAL NETWORK MODELS

Deep neural networks (DNNs) are the foundation of modern AI models and are used in many AI projects, including computer vision, speech recognition and robotics. They are based on the idea that the computation of a neuron involves a weighted sum of input values, where the weighted sum corresponds to the combination of these neuronal values and the value scaling performed by the synapses. Rather than simply outputting a weighted sum, the neuron performs a functional operation on the combined inputs within the neuron. To take advantage of this, researchers need to understand the key design considerations of DNNs and be able to evaluate the usefulness of different DNN design techniques for efficient processing [46].

c: GRAPH-BASED NEURAL NETWORK

In 2005, Gori et al. proposed a new neural network model capable of handling graph data structures: graph neural networks (GNNs). A pioneering study of deep learning methods in non-Euclidean spaces, GNNs generate random state embedding vectors, and the states of nodes are continuously updated according to the information propagation mechanism of the graph. Updates are made based on the state information of neighbouring nodes at a previous time [47]. GNNs are an alternative way to consider additional neighbourhood information in addition to the given data itself.

3) XAI METHODS

It was found that recent papers can be classified into three types of databases, namely clinical [13], [18], [27], [33], registry [10], [12], [20], [24], [25], [29], [34], and knowledge [14], [15], [16], [17], [19], [21], [22], [23], [26], [28], [30], [31], [32], [35]. Additionally, based on review papers of existing XAI models [48], [49], the XAI algorithms used were divided into the following four categories: surrogate

TABLE 2. Publicly available medicine data platform.

Public PV-related Database	Database Description	Data Additional Information	Source	Ref
SIDER	This contains information on marketed medicines and their recorded adverse drug reactions. The available information includes side effect frequency, drug and side effect classifications, as well as links to further information.	5,868 side effects 1,430 drugs	http://sideeffects.embl.de/	[16] [17]
STITCH	This is a database of known and predicted interactions between chemicals and proteins. The interactions include physical and functional associations.	9,643,763 proteins 2,031 organisms	http://stitch.embl.de/	[31]
ChEMBL	This is a manually curated database of bioactive molecules with drug-like properties. It brings together chemical, bioactivity, and genomic data to aid in the translation of genomic information into effective new drugs.	2.3M compounds 43K Indications 14K drugs 1.2K drug warnings	https://www.ebi.ac.uk/chembl/	[21] [31] [35]
DrugBank Online	This is a comprehensive, free-to-access, online database containing information on drugs and drug targets. As both a bioinformatics and a cheminformatics resource, it combines detailed drug data with comprehensive drug target information.	15,234 drug entries (version 5.1.9, released 2022-01-03)	https://go.drugbank.com/	[32] [35]
PubChem	This is an open chemistry database at the National Institutes of Health (NIH). It mostly contains small molecules, but also contains larger molecules such as nucleotides, carbohydrates, lipids, peptides, and chemically-modified macromolecules.	113M compounds 103,988 genes 185,153 proteins	https://pubchem.ncbi.nlm.nih.gov/	[16]
Protein Data Bank	Since 1971, the Protein Data Bank archive (PDB) has served as the sole repository of information about the 3D structures of proteins, nucleic acids, and complex assemblies.	8,165 coordinate entries 7,555 proteins 598 nucleic acids	http://www.wwpdb.org/	[26]
Kaggle	Strongly recommended for beginners in machine learning, this is a great opportunity to practice certain techniques to predict the outcome of the drugs that might be accurate for the patient.	6.03KB drug classification 9.07MB drugs related to medical conditions 11.57MB drugs, side effects and medical condition	https://www.kaggle.com/datasets/prathamtripathi/	[28]
IWPC	A number of pharmacogenomics research centers have collected data sets relating warfarin dosing to a variety of clinical and genetic parameters, including genotypes for CYP2C9 and VKORC1.	640 drugs 100 dosing guidelines 498 drug labels	https://www.pharmgkb.org/page/iwpc	[27]
Linked Western Australian datasets	The WA Data Linkage System (WADLS) is a system of linkages within and between data collections, maintained and operated by the Data Engineering Division.	4,568,496 total number of chains (June 2020)	https://www.datalinkage-wa.org.au/data/available-datasets/	[10]

models, feature importance, attention based, and knowledge distillation and graph based explanations.

a: SURROGATE MODEL

SHapley Additive exPlanations (SHAP) [50] explain individual predictions using the game-theoretic Shapley value [51]. This approach uses the concept of coalitions to compute the Shapley value of features for predicting instances generated by the black-box model. The average marginal contribution of the features in all possible coalitions is the Shapley value.

The main advantage of an explanatory technique such as SHAP is that it has solid roots in game theory, which ensures that the explanation of a prediction instance is fairly distributed across the features. However, SHAP is a slow and computationally expensive explainability technique because it requires Shapley values to be computed for different features in a prediction instance, making it impractical to compute global explanations with multiple prediction instances [52].

The local interpretable model-agnostic explanations (LIME) tool explains the model prediction using the most important contributors. It locally approximates the prediction by perturbing the input around the interest class until a linear approximation is obtained [53] and helps the decision maker to justify the behaviour of the model.

In some studies, there is a model that takes an individual's health information (e.g., medication history) as an input and predicts that an individual will have an adverse outcome from acute coronary syndrome (ACS). XAI was used to detect the contribution of specific drugs to the prediction of ACS. The post-doc use of SHAP and LIME successfully identified both important and relatively less important features [10]. Additionally, there are studies using text analysis to detect adverse drug reactions (ADRs) [15]. In ADR surveillance, the problem of class imbalance often has a negative impact on the outcome. A variant of weighted-CRF was proposed to solve the data imbalance, and LIME was applied to interpret the model results according to either the presence or absence of a weighted loss function.

Layer-wise relevance propagation (LRP) [54] is an explanatory technique applicable to models structured as neural networks, where inputs can be images, videos or text. LRP works by propagating the prediction backwards in the neural network using specifically designed local propagation rules. The propagation procedure implemented by LRP is subject to a conservation property, where what has been received by a neuron must be redistributed to the lower layer in equal amounts. Li et al. [20] proposed a CNN to continuously predict the cytotoxicity of compounds against a leukaemic lymphoblast CCRF-CEM cell line, and applied the LRP technology to a compound that had an important effect on predicting the biological properties of a compound. An attempt was made to identify the important parts of the structure. Compared to previous methods, such as chemical descriptors, it was confirmed that the structural formula of a compound can provide sufficient information for machine learning-based prediction of anticancer efficacy.

b: FEATURE IMPORTANCE

Permutation importance is often used to calculate feature importance: randomly rearrange the values of features and observe the predicted change in the model to determine which features contribute to model predictions. The importance weight is based on the predicted distance between the perturbed and original values of a feature. The importance of a feature can be interpreted as a weight and applied to all features [27]. The importance of a feature is measured by calculating the increase in the model's prediction error after permuting the feature. A feature is considered important if shuffling its values increases the model error because the model relies on that feature for prediction. A feature is considered unimportant if shuffling its values does not change the model error because the model ignores the feature for prediction. In particular, the permutation feature importance measure for random forests was introduced by Breiman and measures the increase in the model's prediction error after permuting feature values to disrupt the relationship between features and the actual outcome [55].

c: ATTENTION MECHANISM

The core concept of the attention mechanism is that the model is designed to pay attention only to the inputs that are most relevant to the predictive task [56]. The attention mechanism has been mainly proposed for natural language processing, including machine translation, and has been found to contribute to improving interpretability as well as technological evolution in the field of visualization [57].

Zhang et al. [12] considered that explainability is important so that stakeholders can trust the deep learning model in pharmacovigilance. Their research provides explainability and transparency by utilizing the attention mechanism for multitasking. Other studies have evaluated more clinically valid explanations for predicting drug side effects [13]. When the attention-based architecture and SHAP (a post-hoc surrogate model) were used together, it enabled appropriate local and global explanations complementary to each other; additionally, the SHAP method was evaluated as a more suitable for a real-time clinical explanation. Liu and Xie [35] developed their novel model, TranSynergy, for a synergistic drug combination prediction. As a knowledge-based and self-attention transformer model, TransSynergy has been demonstrated to improve predictability of synergistic drug combinations as well as improve interpretability. He et al. [21] used traditional machine learning methods such as boosting and Bayesian, as well as five deep learning prediction models including graph attention network and Attentive FP, to discover new drug candidates for the treatment of breast cancer. Attentive FP proposed by Bahdanau et al. [56] is a state-of-the-art graph-based neural network model for predicting molecular properties. It was confirmed that this model can

focus on highly relevant parts of the input by applying the graph attention mechanism and can interpret the learned contents.

d: KNOWLEDGE DISTILLATION & GRAPH BASED EXPLANATIONS

Presently, in the field of deep learning, ''knowledge distillation'' refers to extracting knowledge from a trained model, as seen by the dictionary meaning of ''knowledge'' and ''distillation''. Several studies have been proposed to exploit this extracted knowledge. Knowledge distillation in deep learning can be summarized as a series of processes in which knowledge distilled from a large model (teacher model) is transferred to a small model (student network) [59]. Knowledge extraction is a form of model-specific XAI that extracts knowledge from a large, complex model to a more simplified model, and for many years model compression [60], tree regularization [61], and a combination of model compression and dimensionality reduction [62]. As distillation research has been fragmented, new possibilities have been identified with the development of artificial intelligence interpretability [63]. The rule extraction method, which is similar to knowledge extraction, is a useful XAI method and can be practically applied in the digital medical field [63].

B. ASPECTS OF XAI USE IN PHARMACOVIGILANCE

1) USING XAI TO PREDICT DRUG TREATMENT

The XAI system should be useful for supporting treatment decisions because it suggests a descriptive and visual approach to drug treatment decisions that can help doctors understand and apply recommendations confidently. These results suggest that, even in unguided clinical situations, the XAI system can promote accurate medication prescriptions by doctors [10].

By developing a post-interpretable framework based on an ensemble predictive model to interpret the importance of clinical features and genotypes in predicting daily drug doses, models with fewer variables can be built, and the complexity of the final predictive model can be reduced. Furthermore, it helps clinicians prescribe correct doses to patients using more effective clinical parameters [27]. In recent years, many studies have been conducted to determine an optimal model for accurately predicting drug treatments [10], [21], [27], [29], [30], [31], [32].

For medications associated with Alzheimer's disease, low to moderate doses of antipsychotics, moderate to high doses of antidepressants, and moderate to high doses of cardiovascular drugs have been identified as major features associated with Alzheimer's disease using SHAP [24]. Additionally, an interactive visualization tool has been designed and developed to assist domain users in GNN-based drug repurposing; in this, it was demonstrated how XAI can be used to examine treatments for Alzheimer's Disease (AD) [19]. In the field of cancer research, to gain a deeper understanding of the established models, SHAP was utilized to calculate the contribution of important structural fragments for breast cancer [21]. Additionally, the LRP technique was implemented to interpret the network and visualize the chemical groups predicted by a model that contributes to the toxicity of anticancer drugs with human-readable representations [20].

2) USING XAI TO DETECT ADVERSE DRUG REACTIONS

ADRs are statistically characterized in randomized clinical trials and post-marketing PV; however, their molecular mechanisms remain unknown in most cases. In addition to clinical trials, many elements of knowledge on drug ingredients are available in open-access knowledge graphs, such as their properties, interactions, and involvement in pathways [13]. Rapidly advancing AI technologies can facilitate data-driven estimation when screening multiple variables and capture non-linear relationships to accurately predict clinical outcomes. In recent years, many studies have been conducted to identify an optimal model for accurately predicting ADRs and taking effective preventive measures [17], [18], [25], [26], [27], [28].

3) USING XAI TO EXPLORE POLYPHARMACY

Drug combinations have demonstrated a significant potential for cancer treatment, as they relieve drug resistance and improve therapeutic efficacy. The rapidly increasing number of anticancer drugs has elicited costly and time-consuming experimental investigations of all drug combinations. However, computer technologies may improve the efficiency of drug combination screening. Despite recent advances in applying ML to predict synergistic drug combinations, much room for improvement remains. The XAI method can deconvolute genes contributing to synergistic drug combinations and improve model interpretability [35].

4) EXTRA TASKS: DRUG REPURPOSING, ETC

XAI has been studied for various purposes, such as pharmacodynamics and safety [32], deducing effective properties in the same context [9], drug repurposing [19], and drug classification [13], [28], [33].

IV. OVERALL FINDINGS OF XAI USE IN PHARMACOVIGILANCE

A. BENEFITS OF EXPLAINABLE PHARMACOVIGILANCE

PV XAI can provide clinicians and patients with a more comprehensible support system to ensure interpretability and explainability of AI models [64]. It can also simplify the model evaluation process while increasing the transparency and traceability of the model. There is great potential in systematically monitoring and managing XAI-based models and continuously collaborating with clinicians to enable fine-tuning to advance the system, enabling more efficient and accurate delivery of clinical decision support. The transparency of PV XAI can also manage regulations, risks and other requirements to alleviate PV model governance costs,

minimize manual inspection overhead and costly errors, and reduce the risk of unintentional bias.

B. OPEN PROBLEMS, CHALLENGES, AND NEEDS OF EXPLAINABLE PHARMACOVIGILANCE

XAIs should be developed to enable clinicians to explain decisions in models so that patients can perceive and understand the explanations [65]. The field should consider developing a platform to support communication and linkage with more stable visualizations. This will enable clinicians to support clinical decisions that can build trust with patients and caregivers.

In order to develop a PV XAI model, collaboration with experts in related fields is very important but difficult; however, continuous consideration is required. An appropriate engineering basis for understanding the model used as well as a clinical understanding should be developed.

On the other hand, as models become more and more complex, the flaws of explainability must be accommodated. PV XAI cannot cover all aspects of pharmacovigilance. Therefore, all the life-cycle processes for data-method-result interpretation need to be considered that can balance and reduce bias in the results.

V. CONCLUSION

In this study, we reviewed PV XAI papers and discussed recent research trends and the need for XAI research. Unlike other areas where XAI and AI are developing together, PV XAI research is still in its infancy. There are not many papers on PV XAI and the methodology is limited to a few models. However, studies are slowly beginning to show the potential of XAI research for medication monitoring and patient safety, collecting ADR and ADE information, extracting drug-drug interactions, and predicting drug treatment effects.

As in other areas, as awareness of XAI methods grows, we expect to see AI used in pharmacovigilance and patient safety in many more ways in the coming years than those identified in this review, and the positive potential of XAI for drug therapy, ADRs and interactions is very promising. However, it is clear that the growth of this field may be limited by the lack of validated and established uses of XAI in real-world healthcare settings, and this is an area that requires further investigation. Therefore, the challenges and future prospects of XAIs in pharmacovigilance should be discussed with continued interest.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

ACKNOWLEDGMENT

The authors would like to thank Editage (www.editage.co.kr) for editing and reviewing this manuscript for English language.

REFERENCES

[1] World Health Organization. (Oct. 2004). *WHO Policy Perspectives on Medicines. Looking at the Pharmacovigilance: Ensuring the Safe Use of Medicines*. [Online]. Available: http://apps.who.int/medicinedocs/pdf/s6164e/s6164e.pdf

[2] J. K. Aronson, "Artificial intelligence in pharmacovigilance: An introduction to terms, concepts, applications, and limitations," *Drug Saf.*, vol. 45, no. 5, pp. 407–418, May 2022, doi: 10.1007/s40264-022-01156-5.

[3] R. Ball and P. G. Dal, "'Artificial intelligence' for pharmacovigilance: Ready for prime time?" *Drug Saf.*, vol. 45, no. 5, pp. 429–438, 2022.

[4] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," 2017, *arXiv:1708.08296*.

[5] D. Gunning and D. Aha, "DARPA's explainable artificial intelligence (XAI) program," *AI Mag.*, vol. 40, no. 2, pp. 44–58, 2019, doi: 10.1609/aimag.v40i2.2850.

[6] F. K. Došilovic, M. Brcic, and N. Hlupic, "Explainable artificial intelligence: A survey," in *Proc. 41st Int. Conv. Inf. Commun. Technol., Electron. Microelectron. (MIPRO)*, New York, NY, USA, May 2018, pp. 210–215.

[7] C. Meske and E. Bunde, "Transparency and trust in human-AI-interaction: The role of model-agnostic explanations in computer vision-based decision support," in *Proc. Int. Conf. Hum.-Comput. Interact.* Berlin, Germany: Springer, Jul. 2020, pp. 54–69.

[8] A. Das and P. Rad, "Opportunities and challenges in explainable artificial intelligence (XAI): A survey," 2020, *arXiv:2006.11371*.

[9] M. Tulio, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, no. 1, pp. 1–15.

[10] I. R. Ward, L. Wang, J. Lu, M. Bennamoun, G. Dwivedi, and F. M. Sanfilippo, "Explainable artificial intelligence for pharmacovigilance: What features are important when predicting adverse outcomes?" *Comput. Methods Programs Biomed.*, vol. 212, Nov. 2021, Art. no. 106415, doi: 10.1016/j.cmpb.2021.106415.

[11] *Pharmacovigilance Indicators: A Practical Manual for the Assessment of Pharmacovigilance Systems*, World Health Organization, Geneva, Switzerland, 2015.

[12] S. Zhang, S. Dev, J. Voyles, and A. S. Rao, "Attention-based multi-task learning in pharmacovigilance," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2018, pp. 22324–22328.

[13] J. Rebane, I. Samsten, P. Pantelidis, and P. Papapetrou, "Assessing the clinical validity of attention-based and SHAP temporal explanations for adverse drug event predictions," in *Proc. IEEE 34th Int. Symp. Comput.-Based Med. Syst. (CBMS)*, Jun. 2021, pp. 235–240.

[14] E. Bresso, P. Monnin, C. Bousquet, F.-E. Calvier, N.-C. Ndiaye, N. Petitpain, M. Smaïl-Tabbone, and A. Coulet, "Investigating ADR mechanisms with explainable AI: A feasibility study with knowledge graph mining," *BMC Med. Informat. Decis. Making*, vol. 21, no. 1, pp. 1–14, Dec. 2021.

[15] J. Wang, L.-C. Yu, and X. Zhang, "Explainable detection of adverse drug reaction with imbalanced data distribution," *PLOS Comput. Biol.*, vol. 18, no. 6, Jun. 2022, Art. no. e1010144.

[16] A. S. Mantripragada, S. P. Teja, R. R. Katasani, P. Joshi, and R. Ramesh, "Prediction of adverse drug reactions using drug convolutional neural networks," *J. Bioinf. Comput. Biol.*, vol. 19, no. 1, Feb. 2021, Art. no. 2050046, doi: 10.1142/S0219720020500468.

[17] P. Joshi, M. Vedhanayagam, and R. Ramesh, "An ensembled SVM based approach for predicting adverse drug reactions," *Current Bioinf.*, vol. 16, no. 3, pp. 422–432, Mar. 2021.

[18] Z. Yu, H. Ji, J. Xiao, P. Wei, L. Song, T. Tang, X. Hao, J. Zhang, Q. Qi, Y. Zhou, F. Gao, and Y. Jia, "Predicting adverse drug events in Chinese pediatric inpatients with the associated risk factors: A machine learning study," *Frontiers Pharmacol.*, vol. 12, p. 516, Apr. 2021.

[19] Q. Wang, K. Huang, P. Chandak, M. Zitnik, and N. Gehlenborg, "Extending the nested model for user-centric XAI: A design study on GNN-based drug repurposing," *IEEE Trans. Vis. Comput. Graph.*, vol. 29, no. 1, pp. 1266–1276, Jan. 2023.

[20] Z. Li, Y. W. Lam, Q. Liu, A. Y. K. Lau, H. Y. Au-Yeung, and R. H. M. Chan, "Machine learning-driven drug discovery: Prediction of structure-cytotoxicity correlation leads to identification of potential anti-leukemia compounds," in *Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2020, pp. 5464–5467.

[21] S. He, D. Zhao, Y. Ling, H. Cai, Y. Cai, J. Zhang, and L. Wang, "Machine learning enables accurate and rapid prediction of active molecules against breast cancer cells," *Frontiers Pharmacol.*, vol. 12, pp. 1–15, Dec. 2021, doi: 10.3389/fphar.2021.796534.

[22] J.-B. Lamy, K. Sedki, and R. Tsopra, "Explainable decision support through the learning and visualization of preferences from a formal ontology of antibiotic treatments," *J. Biomed. Informat.*, vol. 104, Apr. 2020, Art. no. 103407.

[23] M. Imran, A. Bhatti, D. M. King, M. Lerch, J. Dietrich, G. Doron, and K. Manlik, "Supervised machine learning-based decision support for signal validation classification," *Drug Saf.*, vol. 45, no. 5, pp. 583–596, May 2022.

[24] B. Braithwaite, J. Paananen, H. Taipale, A. Tanskanen, J. Tiihonen, S. Hartikainen, and A.-M. Tolppanen, "Detection of medications associated with Alzheimer's disease using ensemble methods and cooperative game theory," *Int. J. Med. Informat.*, vol. 141, Sep. 2020, Art. no. 104142.

[25] X. Dong, "An integrated LSTM-HeteroRGNN model for interpretable opioid overdose risk prediction," *Artif. Intell. Med.*, vol. 135, Jan. 2023, Art. no. 102439.

[26] T. Harren, H. Matter, G. Hessler, M. Rarey, and C. Grebner, "Interpretation of structure–activity relationships in real-world drug design data sets using explainable artificial intelligence," *J. Chem. Inf. Model.*, vol. 62, no. 3, pp. 447–462, Feb. 2022, doi: 10.1021/acs.jcim.1c01263.

[27] Y. Zhang, C. Xie, L. Xue, Y. Tao, G. Yue, and B. Jiang, "A post-hoc interpretable ensemble model to feature effect analysis in warfarin dose prediction for Chinese patients," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 2, pp. 840–851, Feb. 2022, doi: 10.1109/JBHI.2021.3092170.

[28] D. V. Gala, "Drug classification using machine learning and interpretability," in *Proc. Smart Technol., Commun. Robot. (STCR)*, Oct. 2021, pp. 1–8.

[29] J. Patel, A. Ladani, N. Sambamoorthi, T. LeMasters, N. Dwibedi, and U. Sambamoorthi, "A machine learning approach to identify predictors of potentially inappropriate non-steroidal anti-inflammatory drugs (NSAIDs) use in older adults with osteoarthritis," *Int. J. Environ. Res. Public Health*, vol. 18, no. 1, p. 155, Dec. 2020, doi: 10.3390/ijerph18010155.

[30] G. Noaro, G. Cappon, G. Sparacino, and A. Facchinetti, "An ensemble learning algorithm based on dynamic voting for targeting the optimal insulin dosage in type 1 diabetes management," in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Nov. 2021, pp. 1828–1831.

[31] A. Lysenko, A. Sharma, K. A. Boroevich, and T. Tsunoda, "An integrative machine learning approach for prediction of toxicity-related drug safety," *Life Sci. Alliance*, vol. 1, no. 6, Dec. 2018, Art. no. e201800098, doi: 10.26508/lsa.201800098.

[32] G. Shtar, L. Rokach, B. Shapira, E. Kohn, M. Berkovitch, and M. Berlin, "Explainable multimodal machine learning model for classifying pregnancy drug safety," *Bioinformatics*, vol. 38, no. 4, pp. 1102–1109, Jan. 2022, doi: 10.1093/bioinformatics/btab769.

[33] I. V. Zhukova, I. V. Derevitskii, G. A. Matveev, T. I. Golikova, and A. Y. Babenko, "Predictive modeling for decision support in the tasks of selecting the drug for obesity treatment," *Proc. Comput. Sci.*, vol. 193, pp. 371–381, Jan. 2021.

[34] Y. Yang, V. Tresp, M. Wunderle, and P. A. Fasching, "Explaining therapy predictions with layer-wise relevance propagation in neural networks," in *Proc. IEEE Int. Conf. Healthcare Informat. (ICHI)*, Jun. 2018, pp. 152–162.

[35] Q. Liu and L. Xie, "TranSynergy: Mechanism-driven interpretable deep neural network for the synergistic prediction and pathway deconvolution of drug combinations," *PLOS Comput. Biol.*, vol. 17, no. 2, Feb. 2021, Art. no. e1008653.

[36] T. Ledieu, G. Bouzillé, F. Thiessard, K. Berquet, P. Van Hille, E. Renault, E. Polard, and M. Cuggia, "Timeline representation of clinical data: Usability and added value for pharmacovigilance," *BMC Med. Informat. Decis. Making*, vol. 18, no. 1, pp. 1–10, Dec. 2018, doi: 10.1186/s12911-018-0667-x.

[37] *The Basics of Clinical Data Registries*. Accessed: Mar. 11, 2021. [Online]. Available: https://www.arbormetrix.com/blog/clinical-data-registry-basics/

[38] E. Pauwels, V. Stoven, and Y. Yamanishi, "Predicting drug side-effect profiles: A chemical fragment-based approach," *BMC Bioinf.*, vol. 12, no. 1, p. 169, Dec. 2011, doi: 10.1186/1471-2105-12-169.

[39] M. Liu, Y. Wu, Y. Chen, J. Sun, Z. Zhao, X.-W. Chen, M. E. Matheny, and H. Xu, "Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs," *J. Amer. Med. Inform. Assoc.*, vol. 19, no. e1, pp. e28–e35, Jun. 2012, doi: 10.1136/amiajnl-2011-000699.

[40] E. Muñoz, V. Nováček, and P.-Y. Vandenbussche, "Facilitating prediction of adverse drug reactions by using knowledge graphs and multi-label learning models," *Briefings Bioinf.*, vol. 20, no. 1, pp. 190–202, Jan. 2019, doi: 10.1093/bib/bbx099.

[41] J. Menche, A. Sharma, M. Kitsak, S. D. Ghiassian, M. Vidal, J. Loscalzo, and A.-L. Barabási, "Uncovering disease-disease relationships through the incomplete interactome," *Science*, vol. 347, no. 6224, 2015, Art. no. 1257601, doi: 10.1126/science.1257601.

[42] D. Szklarczyk, J. H. Morris, H. Cook, M. Kuhn, S. Wyder, M. Simonovic, A. Santos, N. T. Doncheva, A. Roth, P. Bork, L. J. Jensen, and C. von Mering, "The STRING database in 2017: Quality-controlled protein–protein association networks, made broadly accessible," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D362–D368, Jan. 2017, doi: 10.1093/nar/gkw937.

[43] A. Chatr-Aryamontri, B.-J. Breitkreutz, R. Oughtred, L. Boucher, S. Heinicke, D. Chen, and C. Star, "The BioGRID interaction database: 2015 update," *Nucleic Acids Res.*, vol. 43, no. D1, pp. D470–D478, Jan. 2015, doi: 10.1093/nar/gku1204.

[44] H. Kwak, M. Lee, S. Yoon, J. Chang, S. Park, and K. Jung, "Drug-disease graph: predicting adverse drug reaction signals via graph neural network with clinical data," in *Advances in Knowledge Discovery and Data Mining: 24th Pacific–Asia Conference, PAKDD 2020, Singapore, May 11–14, 2020, Proceedings, Part II 24*, Springer, pp. 633–644.

[45] R. Natras, B. Soja, and M. Schmidt, "Ensemble machine learning of random forest, AdaBoost and XGBoost for vertical total electron content forecasting," *Remote Sens.*, vol. 14, no. 15, p. 3547, Jul. 2022.

[46] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, "Efficient processing of deep neural networks: A tutorial and survey," *Proc. IEEE*, vol. 105, no. 12, pp. 2295–2329, Dec. 2017, doi: 10.1109/JPROC.2017.2761740.

[47] M. Gori, G. Monfardini, and F. Scarselli, "A new model for learning in graph domains," in *Proc. IJCNN*, vol. 2, Jul./Aug. 2005, pp. 729–734, doi: 10.1109/IJCNN.2005.1555942.

[48] M. Du, N. Liu, and X. Hu, "Techniques for interpretable machine learning," *Commun. ACM*, vol. 63, no. 1, pp. 68–77, Dec. 2019, doi: 10.1145/3359786.

[49] T. Speith, "A review of taxonomies of explainable artificial intelligence (XAI) methods," in *Proc. ACM Conf. Fairness, Accountability, Transparency*, New York, NY, USA, Jun. 2022, pp. 2239–2250, doi: 10.1145/3531146.3534639.

[50] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," 2017, *arXiv:1705.07874*.

[51] L. Shapley, "A value for N-person games," in *Contributions to the Theory of Games*, vol. 2. Princeton, NJ, USA: Princeton Univ. Press, 1953, pp. 307–317.

[52] K. Främling, M. Westberg, M. Jullum, M. Madhikermi, and A. Malhi, "Comparison of contextual importance and utility with lime and Shapley values," in *Proc. Int. Workshop Explainable, Transparent Auton. Agents Multi-Agent Syst.* Berlin, Germany: Springer, May 2021, pp. 39–54.

[53] 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. Knowl. Discovery Data Mining*, Aug. 2016, pp. 1135–1144.

[54] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS ONE*, vol. 10, no. 7, Jul. 2015, Art. no. e0130140, doi: 10.1371/journal.pone.0130140.

[55] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[56] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.

[57] D. Mascharka, P. Tran, R. Soklaski, and A. Majumdar, "Transparency by design: Closing the gap between performance and interpretability in visual reasoning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4942–4950.

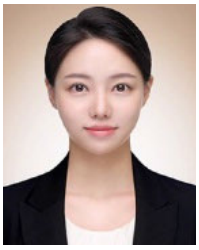
[58] Z. Xiong, D. Wang, X. Liu, F. Zhong, X. Wan, X. Li, Z. Li, X. Luo, K. Chen, H. Jiang, and M. Zheng, "Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism," *J. Medicinal Chem.*, vol. 63, no. 16, pp. 8749–8760, Aug. 2020, doi: 10.1021/acs.jmedchem.9b00959.

[59] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.

[60] A. Polino, R. Pascanu, and D. Alistarh, "Model compression via distillation and quantization," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–16. [Online]. Available: https://openreview.net/forum?id=S1XolQbRW
[61] M. Wu, M. Hughes, S. Parbhoo, M. Zazzi, V. Roth, and F. Doshi-Velez, "Beyond sparsity: Tree regularization of deep models for interpretability," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, 2018, pp. 1–15.
[62] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," *Electronics*, vol. 8, no. 8, p. 832, 2019.
[63] G. Yang, Q. Ye, and J. Xia, "Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond," *Inf. Fusion*, vol. 77, pp. 29–52, Jan. 2022.
[64] G. Vilone and L. Longo, "Notions of explainability and evaluation approaches for explainable artificial intelligence," *Inf. Fusion*, vol. 76, pp. 89–106, Dec. 2021.
[65] T. A. J. Schoonderwoerd, W. Jorritsma, M. A. Neerincx, and K. van den Bosch, "Human-centered XAI: Developing design patterns for explanations of clinical decision support systems," *Int. J. Hum.-Comput. Stud.*, vol. 154, Oct. 2021, Art. no. 102684.

SEUNGHEE LEE received the B.S. degree in mathematics and the Ph.D. degree in dynamical systems from Chungnam National University (CNU), in 2004 and 2013, respectively. She was a Postdoctoral Researcher with the MathVision2020 Project, CNU, from 2013 to 2016, and the Medical Data Analytics Team, National Institute for Mathematical Sciences (NIMS), from 2016 to 2019. She is currently a Research Professor with the Healthcare Data Science Center, Konyang University Hospital, Daejeon, South Korea. Her research interests include dynamical systems, mathematical data analytics, and medical AI.

SEONYOUNG KIM received the B.S. degree in information statistics and the M.S degree in bio-AI convergence from CNU, in 2019. She is currently a Data Scientist with the Healthcare Data Science Center, Konyang University Hospital, Daejeon, South Korea. Her research interests include statistical data analytics, biological AI, medical AI, and explainable AI.



JIEUN LEE received the degree from Hannam University. She is currently a Data Scientist with the Healthcare Data Science Center, Konyang University Hospital, Daejeon, South Korea. Her research interests include big data analysis and artificial intelligence in the medical field.

JONG-YEUP KIM received the M.D., M.S., and Ph.D. degrees from the College of Medicine, Konyang University, Daejeon, South Korea, in 2003, 2006, and 2018, respectively. He is currently the Head of the Healthcare Data Science Center, Konyang University Hospital, a Head Professor with the Department of Information Medicine, and an Associate Professor with the Department of Otolaryngology, College of Medicine, Konyang University. His current research interests include health and medical big data, medical artificial intelligence, and clinical decision support systems.

MI-HWA SONG received the B.S. degree from the Department of Computer Science, Ewha Womans University, in 2002, the M.S. degree from the Department of Computer Science and Engineering, Seoul National University, in 2005, and the Ph.D. degree from the Department of Computer Science, South Korea, in 2010. She was a Visiting Scholar with the Department of Radiology, Imaging Science and Information Systems (ISIS) Center, Georgetown University, USA, worked on a proposal for building fault-tolerant medical information systems for the DoD. She is currently a Professor with Semyung University, South Korea. Her research interests include computational intelligence, big data analysis, and digital therapeutics.

SUEHYUN LEE received the B.S. degree in biomedical engineering from Gachon University, in 2010, and the M.S. and Ph.D. degrees in medical informatics from Seoul National University, in 2012 and 2017, respectively. She is currently an Assistant Professor with the College of IT Convergence, Gachon, Seongnam, South Korea. Her research interests include pharmacovigilance and PGHD.