DOCTORAL THESIS

A Novel Explainable Artificial Intelligence Framework with Improved Learning Mechanisms

Sami Kabir

Pervasive and Mobile Computing (PMC)

LULEÅ UNIVERSITY OF TECHNOLOGY

A Novel Explainable Artificial Intelligence Framework with Improved Learning Mechanisms

Sami Kabir

Pervasive and Mobile Computing
Department of Computer Science, Electrical and Space Engineering
Luleå University of Technology
Skellefteå, Sweden

Supervisors:

Prof. Karl Andersson and Prof. Mohammad Shahadat Hossain

To my parents, spouse, and daughter...

Abstract

Artificial Intelligence (AI) models are increasingly being deployed in various critical domains, such as, healthcare, finance, law, and autonomous systems, resulting in remarkable benefits for the society. Due to black-box (sub-symbolic) nature, these AI models do not provide explanation of the predictive output. This lack of explanation results in lack of transparency between human and machine. Such opacity of black-box AI models, particularly deep learning architectures, is a significant concern. To address this concern, Explainable Artificial Intelligence (XAI) has emerged as a vital research area. The aim of XAI is to enhance transparency, trust, and accountability of AI models. Various post-hoc XAI tools explain the outputs of black-box AI models based on training datasets. However, the use of training datasets, rather than domain knowledge, makes such explanations proxy. A biased training dataset may lead to a misleading post-hoc explanation as well. In contrast, an explanation based on domain knowledge will be more trustworthy to an end user. Motivated by this, we propose a novel XAI framework, consisting of Belief Rule Based Expert System (BRBES), to predict output and explain it with reference to domain knowledge. BRBES represents domain knowledge with its rule base. It outperforms other knowledge-driven AI models to handle uncertainty due to ignorance. To improve the accuracy of BRBES, we fine-tune its parameters and structure through evolutionary learning. Moreover, to overcome the scarcity of labeled dataset in this learning mechanism, we integrate semi-supervised and self-supervised learning with BRBES. We explain the output of BRBES with respect to the most influential rule of the rule base of BRBES. Thus, the output of our proposed XAI framework becomes not only accurate, but also explainable.

This doctoral thesis delves into the challenges and opportunities surrounding XAI in order to provide a comprehensive understanding of the AI output. It presents a novel XAI framework to provide domain knowledge based energy consumption prediction with improved accuracy, and explain this predictive output in non-technical human language. Then it deals with multi-modal air quality data by integrating deep learning model with BRBES. Moreover, to reduce dependence on labeled data for evolutionary learning, this thesis integrates semi-supervised and self-supervised learning with BRBES.

This thesis presents six significant contributions. First, we conduct a Systematic Literature Review (SLR) on XAI using the PRISMA guidelines, delving into the numerous challenges and opportunities of XAI. Extensive research is conducted to explore the definition, terminologies, taxonomy, and application domains of XAI. It highlights various challenges of XAI, such as, no universal definition, trade-off between accuracy and explainability, and lack of standardized evaluation metrics. To address this lack of

standardized evaluation metrics, we also propose a unified framework to evaluate XAI. Secondly, we present an innovative explainable BRBES (eBRBES) framework, which offers accurate prediction of building energy consumption phenomenon, while providing insightful explanation and counterfactual based on domain knowledge. As part of eBRBES framework, we also present a novel Belief Rule Based adaptive Balance Determination (BRBaBD) algorithm to assess the optimal balance of the proposed framework between explainability and accuracy. Thirdly, we propose a mathematical model to integrate BRBES with the Convolutional Neural Network (CNN). We leverage the domain knowledge of BRBES, and image data patterns discovered by CNN with this integrated approach. We predict air quality with this integrated model using outdoor ground images and numerical sensor data. Fourth, we integrate two-layer BRBES with CNN to monitor air quality using satellite images, and relevant environmental parameters, such as, cloud, relative humidity, temperature, and wind speed. The two-layer BRBES showcases the strength of BRBES in conducting reasoning across multiple layers. Fifth, we enhance the learning mechanism of BRBES by utilizing the extensive unlabeled energy data, along with limited labeled data. For this purpose, we synthetically generate unlabeled data through weak and strong augmentation. We then pseudo-label these unlabeled data by integrating semi-supervised self-training model with BRBES. By learning from both labeled and pseudo-labeled data, initial BRBES transitions to semi-supervised BRBES. Sixth, to conduct the learning mechanism of BRBES without relying on any labeled data, we integrate self-supervised learning with it. To this effect, we pseudo-label the synthetically generated unlabeled energy data with BRBES in the pretext tasks of self-supervised learning. Then, the initial BRBES learns deeper representation from these pseudo-labeled data, and transitions to self-supervised BRBES. We then transfer this BRBES, learned in a self-supervised approach, to the downstream task for performing regression of energy consumption.

Based on the research findings of this thesis applied on two different phenomena, it can be argued that the proposed XAI framework provides prediction with greater precision, and explanation with higher interpretability.

Contents

Acknowledgments

I would like to thank my supervisors Prof. Dr. Karl Andersson, and Prof. Dr. Mohammad Shahadat Hossain for providing me the opportunity to conduct my doctoral research at Luleå University of Technology (LTU), Skellefteå, Sweden. Both of them supervised my doctoral research scientifically and coherently. I am grateful to Prof. Dr. Christer Åhlund and Dr. Karan Mitra for their support and encouragement during this research. I deeply appreciate my fellow doctoral students and the other staff members of LTU for the supportive environment.

I would like to express my gratitude to Rönnbäret Foundation, Sweden and VINNOVA (Sweden's Innovation Agency) for supporting this study through the Digital Stadsutveckling Campus Skellefteå project, under grant number 2022-01188. I am also grateful to Patrik Sundberg of Skellefteå Kraft for providing me the hourly energy consumption dataset of residential apartments of Skellefteå city, Sweden.

Finally, I would like to thank my parents, spouse, daughter, and other family members for their sacrifices to make my research journey a success. Their presence in my life makes me feel proud.

Skellefteå, Sweden, November 2025
Sami Kabir

Part I: Thesis Introduction

Chapter 1

Introduction

1.1 Introduction

After mechanization, electrification, and digitalization, the fourth stage of industrialization has been termed as "Industry 4.0". The concept of industry 4.0 puts emphasis on the integration of information technology with industrial technology, resulting in more efficient mass production. Internet of Things (IoT), big data analytics, and Artificial Intelligence (AI) are vital components of industry 4.0 [1]. IoT is a global network of objects, which captures, stores, and processes sensor data through embedded systems [2]. Such embedded systems are constituted by Radio Frequency IDentification (RFID) and Wireless Sensor Network (WSN) technologies [3]. Data coming from various IoT sensors, and other sources constitute big data [4]. By applying AI based analytics on big data, meaningful prediction can be computed by discovering hidden data pattern. Prediction can be computed in two ways: data-driven approach, and knowledge-driven approach [5]. Traditional AI models, such as, machine learning and deep learning fall under the data-driven approach. Machine learning builds a statistical model of the training data without being explicitly programmed to do so. It does not have feature extractor to convert natural raw data into feature vector [6]. On the other hand, deep learning, based on neural network architecture, discovers hidden features of raw data by applying representation learning method. These conventional AI models can predict accurately because of their intricate mathematical operations. Such complex mathematical equations are not explainable in human language to a human user. Hence, predictive outputs of machine learning and deep learning models are non-transparent to a human user. This lack of transparency hampers the relationship of trust between human and AI.

However, human trust is a prerequisite for widespread adoption of AI in society, which has triggered the emergence of Explainable Artificial Intelligence (XAI) as a vital research area [7] [8]. This trust is more significant for safety-critical application areas, such as, healthcare, autonomous vehicles, and energy domain. Various post-hoc explanation techniques, such as, feature relevance, text explanation, and visual explanation

are used to explain the output of a machine learning or deep learning model [9]. Such post-hoc explanation is based on training dataset. It does not consider knowledge or facts of the concerned domain or application area. Hence, post-hoc explanation is proxy. Biased or erroneous training data will lead to misleading post-hoc explanation [10]. On the contrary, knowledge-driven approach is constituted by expert system, which has two parts: knowledge base and inference engine [11]. Knowledge base is constituted by if-then rules, which represents knowledge or facts of the relevant domain. Inference engine conducts reasoning over input data against this knowledge base to predict output. Explanation made by a knowledge-driven approach with respect to its knowledge base is more trustworthy to a user compared with post-hoc explanation of data-driven approach [10]. Fact based explanation, instead of training data based explanation, is the key reason of this human trust in knowledge-driven approach. Hence, we argue that knowledge-driven approach outperforms data-driven approach in terms of trustworthiness of explanation. Belief Rule Based Expert System (BRBES) [11], fuzzy logic, MYCIN [12], and PERFEX [13] are some of the examples of knowledge-driven approach. However, human error or ignorance may cause input data to become wrong or blank. Such input data result in inaccurate predictive output, leading to uncertainties [14]. These uncertainties are caused by imprecise or incomplete human knowledge [15][16]. Low battery power, computational and memory constraints, inadequate communication bandwidth, malicious attacks, and harsh environments are also the factors which cause missing, duplicate, or inconsistent sensor data, resulting in uncertainties [14] [17]. Such uncertainties due to ambiguity are captured by the distributed belief structure of the rule base of BRBES [18]. Moreover, to deal with uncertainty due to ignorance arising from missing input data, BRBES updates the belief degrees of its rule base [11]. Thus, BRBES outperforms other knowledge-driven approaches to deal with uncertainties, specially those caused by ignorance due to input data unavailability [11]. BRBES, being a rule-based model, falls under ante-hoc (intrinsic) XAI method. Therefore, this thesis proposes a novel XAI framework, consisting of BRBES, to predict and explain output based on domain knowledge, while handling data uncertainties [10]. However, BRBES, being a knowledge-driven approach, has lower accuracy compared with traditional AI models [7]. An explanation, despite being domain knowledge based, will not be reliable to a user if it has lower accuracy. To improve the predictive accuracy of BRBES, we fine-tune its parameters and structure with Joint Optimization of Parameter and Structure (JOPS) algorithm [19]. JOPS requires supervised labeled data to optimize BRBES. However, in many cases, availability of such labeled data becomes either expensive or difficult due to privacy protection. To overcome the scarcity of balanced labeled data, this thesis integrates semi-supervised learning with BRBES [20]. To this effect, BRBES utilizes initial labeled data to optimize its parameters and structure through JOPS until its accuracy reaches the confidence threshold. Next, unlabeled data are generated through weak and strong augmentation. As semi-supervised learning, self-training model is employed to pseudo-label these unlabeled data with BRBES. These labeled and pseudo-labeled data are then utilized to optimize BRBES, resulting in a semi-supervised BRBES. This semi-supervised BRBES optimizes itself with limited labeled data before pseudo-labeling the unlabeled data. Hence, to optimize BRBES without any

labeled data, we integrate self-supervised learning with BRBES [21]. In the pretext tasks of self-supervised learning, we employ BRBES to pseudo-label the synthetically generated unlabeled data. Then the initial BRBES optimizes itself with these pseudo-labeled data to transition to the self-supervised BRBES. We then transfer this self-supervised BRBES to the downstream task for performing regression. To make the predictive output of BRBES trustworthy, we explain the rationale in support of this output to the end user in natural human language. We also provide counterfactual to let the user understand how an opposite output could have been achieved. To evaluate the balance of this BRBES framework between explainability and accuracy, we propose Belief Rule-Based adaptive Balance Determination (BRBaBD) algorithm [10]. As application area of this novel XAI framework, this thesis predicts energy consumption of buildings.

Moreover, we make this framework more robust by dealing with multi-modal input data. To this effect, this thesis proposes a mathematical model to integrate BRBES with Convolutional Neural Network (CNN) for dealing with image data [22]. In this integrated model, CNN predicts output from image data, which is then fed to BRBES as input. Thus, this integrated model handles image data with CNN, and numerical data with BRBES, leading to a more robust final output. This thesis takes air pollution prediction as an application area of this integrated model, where outdoor images and numerical sensor data are provided as input to CNN and BRBES respectively. As part of this integration with CNN, this thesis also proposes two-layer BRBES, which demonstrates the strength of BRBES to conduct reasoning across multiple layers [23]. With this integrated model of CNN and two-layer BRBES, we monitor air quality using satellite images, and relevant environmental parameters, such as, cloud, relative humidity, temperature, and wind speed. Thus, this thesis provides a novel XAI framework, consisting of BRBES, to provide domain knowledge based accurate prediction and explanation with reduced dependence on supervised labeled data. Motivation of this thesis is presented in the next sub-section.

1.2 Motivation of Thesis

Industry 4.0 has been triggered by big data, high computing power, and predictive analytics of AI models. AI has proliferated in every sphere of our life in society, ranging from product recommendation during online shopping to interaction with chatbots. Computationally complex AI models can predict phenomenon with high accuracy [24]. Deep learning, a highly accurate AI model, adjusts the values of its weights by learning from training data. However, it is extremely tough to link these numerical weights with physical context of the problem. This makes explanation of a deep learning outcome to a human user in human language very complex. However, when a decision by an AI model impacts human lives, such as, disease prediction, treatment recommendation, and self-driven vehicles, knowing the rationale behind this decision becomes significant [7]. If training dataset of a medical AI does not include certain population group, this AI will provide erroneous prediction when applied on that unrepresented group. Hence, inclusive and unbiased dataset is a major precondition to train AI models properly [25]. Artificial

Intelligence Act, passed by the European Parliament, also puts emphasis on the explanation of a decision made by an AI model [26]. Due to rising application of AI models in the society, European Union (EU) has introduced "right to explanation" in General Data Protection Regulation (GDPR) [27]. This regulation entitles any EU citizen to demand explanation of a decision made by an AI model, which directly impacts his/her personal life. In general, humans are unwilling to trust an opaque AI model's uninterpretable output [28].

To ensure safe adoption of AI models in human lives, various post-hoc tools, such as, Local Interpretable Model-Agnostic Explanations (LIME) [29], SHapley Additive exPlanation (SHAP) [30], text explanation, and visualization are employed to explain black-box AI models. However, such post-hoc tools provide explanation based on training dataset. Wrong, biased, or non-inclusive training dataset will result in erroneous explanation [10]. Fact or knowledge of the concerned application area is not taken into account by these post-hoc tools. Hence, post-hoc explanations are proxy [10]. Contextual Importance and Utility (CIU) is a model-agnostic XAI method, which produces instance-specific explanations of black-box AI models [31]. CIU provides explanations without any intermediate interpretable model [32]. Here, feature importance expresses the impact of changing the value of a feature on the model outcome compared to other features. Utility provides instance-level assessment of how favorable a feature value is to the outcome. However, CIU explanation does not consider knowledge of the relevant application area. To overcome the limitation of such proxy explanation of data-driven AI models, this thesis proposes a knowledge-driven AI model [10]. Explanation of a knowledge-driven AI model reflects relevant domain knowledge, instead of training dataset. Thus, knowledge-driven AI makes factual explanation, rather than proxy explanation of data-driven AI. Such factual explanation gains more trust of a human user than proxy explanation [10].

Uncertainties associated with input data is a severe factor which hampers the prediction accuracy of any AI model. Such uncertainties are triggered by wrong or blank input data, which occurs due to imprecise or incomplete human knowledge [22]. Technical glitch of a device may also cause such uncertainties [22]. Addressing such uncertainties due to ambiguity and ignorance is of paramount importance to improve the prediction accuracy of an AI model. BRBES, a knowledge-driven AI model, outperforms other knowledge-driven approaches to deal with uncertainties, specially due to ignorance [18]. BRBES handles such uncertainties with belief rule base, and belief degree update mechanism. To enhance the predictive accuracy of BRBES further, we optimize its parameters and structure with JOPS algorithm. However, JOPS requires historical labeled data to perform this optimization. Access to such labeled data may become challenging due to privacy concern or costly due to subscription fee. To overcome this shortage of labeled data, this thesis integrates semi-supervised [20] and self-supervised [21] learning with BRBES. To make BRBES more robust by handling multi-modal input data, we integrate it with deep learning model [22]. We also employ two-layer BRBES for more in-depth reasoning over a higher number of input data [23]. Thus, BRBES provides not only factual explanation, but also accurate prediction.

To gain human trust, an AI model has to predict with accuracy, and explain with

domain knowledge. A balance between explainability and accuracy is critical for safe application of an AI model. An AI model with high explainability but low accuracy or vice versa will not be reliable. Hence, this thesis proposes BRBaBD algorithm to evaluate the balance of an AI model between explainability and accuracy [10]. Motivated by all of the aforementioned factors, this thesis proposes a novel XAI framework, consisting of BRBES, to provide domain knowledge based accurate prediction and explanation, while addressing data uncertainties and shortage of labeled data. To act on this motivation, this thesis identifies three research questions. We present these questions in the next sub-section.

1.3 Research Questions

This sub-section presents the research questions based on state-of-the-practice and state-of-the-art review. Three main thesis contributions are: 1) a novel XAI framework to provide domain knowledge based explanation of predictive output under uncertainty [10], 2) a mathematical model to integrate knowledge-driven approach (BRBES) with data-driven approach (deep learning) for improved accuracy and robustness [22] [23], and 3) an improved learning mechanism of BRBES to overcome the scarcity of labeled data [20] [21]. In the following, we describe three research questions, which this thesis answers:

RQ1: What opportunities exist to explain predictive output with domain knowledge while addressing explainability versus accuracy balance?

An explanation based on knowledge of the concerned domain, instead of post-hoc training dataset, is of paramount importance to gain trust of human users. Hence, this thesis proposes a novel XAI framework, incorporating BRBES, to provide domain knowledge based explanation [10]. BRBES represents domain knowledge with its rule base, while handling data uncertainties. Domain knowledge based explanation and counterfactual are communicated to the user in human language through an interface. To address the explainability versus accuracy balance of this framework, this thesis proposes a novel BRBaBD algorithm [10]. Thus, **RQ1** is answered by this thesis.

RQ2: How to improve accuracy and robustness of BRBES while handling data uncertainties?

Accuracy of predictive output is very significant to enhance safety and reliability of an AI model. BRBES has lower accuracy than complex black-box AI models. Hence, this thesis improves accuracy of BRBES through a learning mechanism. This mechanism optimizes the parameters and structure of BRBES through an evolutionary algorithm JOPS. Moreover, this thesis makes BRBES more robust by dealing with multi-modal input data. For this purpose, it proposes a mathematical model to integrate CNN with BRBES [22]. This integrated model processes image data with CNN, and numerical data with BRBES. As part of this integration, this model also employs two-layer BRBES to perform reasoning over a higher number of input features [23]. BRBES deals with uncer-

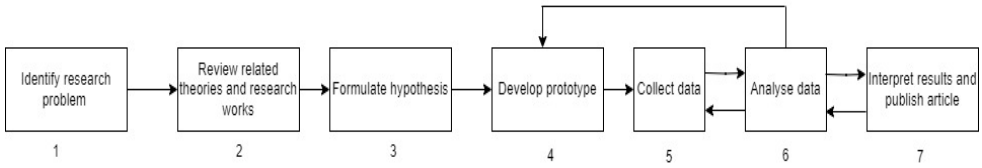

Figure 1.1: Research process

tainties due to ambiguity and ignorance with its belief rule base and belief degree update mechanism respectively. Thus, this thesis answers **RQ2**.

RQ3: What are the methods overcoming the shortage of labeled data in the learning mechanism of BRBES?

Learning mechanism improves accuracy of BRBES by optimizing its parameters and structure with JOPS algorithm. JOPS learns representation from supervised labeled data to perform this optimization. However, such labeled data become scarce due to cost and privacy issues. Hence, shortage of such labeled data has to be overcome to uphold the learning mechanism of BRBES. To this effect, this thesis integrates semi-supervised [20] and self-supervised learning [21] methods with BRBES in its XAI framework. Semi-supervised BRBES learns from limited labeled data before optimizing itself with pseudo-labeled data. On the other hand, self-supervised BRBES does not rely on any labeled data. It pseudo-labels the unlabeled data in the pretext tasks, learns representation from such pseudo-labeled data, and transfers this learning to downstream task to provide prediction. Thus, this thesis answers **RQ3**.

To answer the aforementioned three research questions, this thesis performs systematic investigation through research methodology. This research methodology is outlined in the next sub-section.

1.4 Research Methodology

This sub-section presents the research methodology followed in this thesis work. The research methodology is defined as a process to systematically address a research problem [33]. The methodology consists of interviews, surveys and other data collection techniques, and covers both present and historical information. This thesis work follows theoretical research. As novel theoretical contribution, this thesis makes an AI model explainable and accurate, while overcoming the scarcity of labeled data in the learning mechanism. To conduct this study, the following steps have been followed: 1) Identification of the research problem; 2) Investigation of the related theories, and previous research works; 3) Formulation of hypothesis to address research gap; 4) Development of a prototype software; 5) Data collection and testing with the research prototype; 6) Analysis of the research results; and 7) Interpretation and reporting of research results

in a research article. This research process is shown in Figure 1.1. We repeat steps 4 to 6 until the accuracy and explainability of our framework reaches an acceptable level.

As an example, the research process for answering the third research question is described below.

Step 1 The research problem is the lack of labeled data to conduct the learning mechanism of BRBES, which is described in sub-section 1.3.

Step 2 Related theories and previous research on addressing the shortage of labeled data have been reviewed. Important related works on semi-supervised learning have been mentioned in sub-section 2.2.4.

Step 3 Based on the findings from Step 2, a hypothesis to integrate semi-supervised learning with BRBES was developed, which is described in Paper E. A novel semi-supervised explainable BRBES framework was proposed to pseudo-label the augmented unlabeled data, and learn from both labeled and pseudo-labeled data.

Step 4 A prototype software was developed using python and C++ based on the hypothesis, which is presented in Paper E.

Step 5 To evaluate the semi-supervised explainable BRBES framework, hourly energy consumption data of 62 residential apartments of Skellefteå, Sweden were collected from Skellefteå Kraft of Sweden. In addition to this labeled dataset, this framework synthetically generated unlabeled data through augmentation. These unlabeled data were then pseudo-labeled by BRBES. These pseudo-labeled data were utilized to overcome the shortage of labeled data in the learning mechanism of BRBES.

Step 6 The results were analyzed and the performance of the framework was evaluated using two accuracy metrics, five explainability metrics, and two counterfactual metrics against four state-of-the-art AI models.

Step 7 After validation of the hypothesis, the results and findings were documented and published in a peer reviewed journal.

For this third research question, as part of evolutionary algorithm, steps 4 to 6 were repeated multiple times to achieve better results. The above research process was followed to address the first and second research questions throughout this thesis as well. Next sub-section highlights the contributions of this thesis with respect to each of the three research questions.

1.5 Thesis Contributions

In this sub-section, we present the research contributions of this thesis. A mapping between the thesis contributions and research questions are shown in Table 1.1. This illustrates a high-level overview of the way our research findings address the stated research questions from sub-section 1.3. The main contributions include:

1. performing a literature review of XAI from the perspectives of challenges, and opportunities (RQ1),

2. proposing an advanced explainable belief rule-based framework to accurately predict the energy consumption of buildings, and explain it with domain knowledge (RQ1,2),

3. proposing an integrated approach of belief rule base and deep learning to predict air pollution from ground images and sensor data (RQ2),

4. proposing an integrated approach of two-layer belief rule base, and CNN to monitor air quality from satellite images, and environmental data (RQ2),

5. proposing a semi-supervised explainable BRBES framework to improve learning mechanism of BRBES by learning from limited labeled data, and utilizing extensive unlabeled data (RQ1,2,3), and

6. proposing a self-supervised explainable BRBES framework, where learning mechanism improves accuracy of BRBES without requiring any labeled data (RQ1,2,3).

In the following, we list the produced research results, and discuss the individual contributions in detail.

- **Paper A.** *A Review of Explainable Artificial Intelligence from the Perspectives of Challenges and Opportunities.* **Sami Kabir**, Mohammad Shahadat Hossain, and Karl Andersson. MDPI Algorithms. Published on September 3, 2025.

– The increasing use of Artificial Intelligence (AI) models across various critical domains, such as, healthcare, finance, law, and autonomous systems has delivered remarkable societal benefits. However, due to the black-box (sub-symbolic) nature, these AI models provide prediction without explaining the underlying reasons to the end user. This lack of explanation causes lack of transparency between humans and machines. Such opacity of complex AI models, specially deep learning models, is a concern for safe application of AI models. To address this concern, Explainable Artificial Intelligence (XAI) has emerged as a pivotal area of research. Enhancing transparency, trustworthiness, and accountability of AI models is the aim of XAI. This review article presents a comprehensive overview of XAI by covering its foundational concepts, definitions, terminologies, taxonomy, and application domains.

Table 1.1: Mapping between paper contributions and research questions

Paper	Title	RQ1	RQ2	RQ3
A	A Review of Explainable Artificial Intelligence from the Perspectives of Challenges and Opportunities [7]	X		
B	An Advanced Explainable Belief Rule-Based Framework to Predict the Energy Consumption of Buildings [10]	X	X	
C	An Integrated Approach of Belief Rule Base and Deep Learning to Predict Air Pollution [22]		X	
D	An integrated approach of Belief Rule Base and Convolutional Neural Network to monitor air quality in Shanghai [23]		X	
E	A Semi-Supervised-Learning-Aided Explainable Belief Rule-Based Approach to Predict the Energy Consumption of Buildings [20]	X	X	X
F	A Self-Supervised Belief Rule-Based Approach to Predict and Explain the Energy Consumption of Buildings [21]	X	X	X

This article also highlights various challenges of XAI, such as, no universal definition, trade off between accuracy and interpretability, and lack of standardized evaluation metrics. To direct future XAI research, this paper puts emphasis on human-centric design, interactive explanation, and standardized evaluation frameworks. This review paper serves as a resource for researchers, practitioners, and policymakers to explore the explainability of AI models.

– I was the main contributor of this XAI review paper, which provides a comprehensive review with respect to the first research question. I employed Systematic Literature Review (SLR) [34] approach using the PRISMA guidelines [35] to identify reliable XAI studies published during 2022 to 2025. Based on the findings of this literature review, I wrote the complete review paper focusing on the state-of-art XAI, along with its challenges, and opportunities. During the process, my supervisors supervised my work through sprint meetings. Their constructive comments helped me to improve the content, and structure of the review article.

- **Paper B.** *An Advanced Explainable Belief Rule-Based Framework to Predict the Energy Consumption of Buildings.* **Sami Kabir**, Mohammad Shahadat Hossain, and Karl Andersson. MDPI Energies. Published on April 9, 2024.

– Providing predictive energy consumption of buildings to utility companies, users, and facility managers can contribute to reduce energy waste. However, application of such prediction in this domain is limited due to various drawbacks of prediction

algorithms, such as, non-transparent output, ad hoc explanation by post hoc tools, low accuracy, and the inability to deal with data uncertainties. Hence, accurate prediction with domain knowledge-based explainability is of paramount importance to make energy prediction trustworthy. Motivated by this, we propose an advanced explainable Belief Rule-Based Expert System (eBRBES) framework to predict energy consumption accurately, and explain it with domain knowledge. We improve prediction accuracy of BRBES by optimizing its parameters and structure, while handling data uncertainties using its inference engine. To predict energy consumption, we consider floor area, daylight, indoor occupancy, and building heating method as influential parameters. In addition to explanation, we also communicate how a counterfactual output on energy consumption could have been achieved. Both explanation and counterfactual are communicated through an user interface. Furthermore, we propose a novel Belief Rule-Based adaptive Balance Determination (BRBaBD) algorithm to evaluate the balance of our model between explainability and accuracy. To validate the proposed eBRBES framework, we use building energy consumption dataset of Skellefteå, Sweden. BRBaBD results show that our proposed eBRBES framework is 85.08% more balanced than state-of-the-art machine learning algorithms.

– I was the main contributor of this paper, which covers both first and second research questions by providing domain knowledge based accurate prediction and explanation. My supervisors provided me with insightful guidance to develop the proposed eBRBES framework, and identify the relevant metrics to evaluate explainability and counterfactual. This paper assessed the balance of AI models between explainability and accuracy with BRBaBD algorithm. We validated the proposed eBRBES framework with the hourly energy consumption dataset of 62 residential apartments of Skellefteå, Sweden. This dataset was provided by Skellefteå Kraft of Sweden under the VINNOVA (Sweden's Innovation Agency) funded Digital Stadsutveckling Campus Skellefteå project.

- **Paper C.** *An Integrated Approach of Belief Rule Base and Deep Learning to Predict Air Pollution.* **Sami Kabir**, Raihan Ul Islam, Mohammad Shahadat Hossain, and Karl Andersson. MDPI Sensors. Published on March 31, 2020.

– Internet of Things (IoT) has triggered widespread availability of sensor data. Such sensor data are used by predictive algorithms to predict phenomena by applying reasoning. Issuing public health warning based on air pollution prediction, and reducing building energy waste based on energy consumption prediction are some of the use cases of sensor data stream where accurate prediction is critical to protect people and assets. However, prediction accuracy is severely hampered by uncertainties of sensor data. Such uncertainties are handled by knowledge base and inference engine of Belief Rule Based Expert System (BRBES). This is a knowledge-driven

approach, which can handle uncertainties better than other knowledge-driven techniques, such as, fuzzy logic and Bayesian probability theory. On the other hand, deep learning, a data-driven technique, learns hidden representation of data by applying analytics. Thus, deep learning infers prediction by reasoning over available data stream. Combined application of BRBES and Deep Learning can result in more accurate prediction by addressing sensor data uncertainties while discovering hidden data pattern. Hence, this paper proposes a novel mathematical model to integrate BRBES with deep learning to capture the nonlinear dependencies among relevant variables. We improved the accuracy of BRBES further by fine-tuning its parameters and structure. We take air pollution prediction as application area of our proposed combined approach. This model has been evaluated against both synthetic and real-world dataset, consisting of outdoor images and $PM_{2.5}$ concentrations. Our proposed model also distinguished a hazy image between polluted air and fog by calculating the difference between temperature and dew point. According to the experimental results, the proposed integrated model demonstrated higher prediction accuracy than only BRBES and only Deep Learning.

– I was the main contributor to the paper, which covers the second research question. For this research, at first, we employed CNN on outdoor ground images to evaluate the level of air pollution in terms of $PM_{2.5}$. Then we employed BRBES, which takes CNN prediction and $PM_{2.5}$ sensor data as its two input values. Finally, BRBES predicted the air pollution in terms of Air Quality Index (AQI). Thus, the proposed integrated approach of this paper dealt with multi-modal input data (both image data and numerical sensor data), contributing to the robustness of BRBES. We applied our proposed predictive algorithm on two different datasets. One is a labeled dataset of synthetic images with corresponding $PM_{2.5}$ label. The other one consists of hourly real images of China's Shanghai city. These images were paired with $PM_{2.5}$ concentrations, temperature, dew point, and relative humidity data of the same place and same time. My supervisors took part in the complete process of developing the proposed model, and provided effective feedback. Also, they performed an iterative proofreading process to improve the content and structure of the paper.

- **Paper D.** *An integrated approach of Belief Rule Base and Convolutional Neural Network to monitor air quality in Shanghai.* **Sami Kabir**, Raihan Ul Islam, Mohammad Shahadat Hossain, and Karl Andersson. Elsevier Expert Systems With Applications. Published on June 20, 2022.

– Accurate monitoring of air quality can contribute to protect public health. Fine particulate matter ($PM_{2.5}$) concentrations and ground images can be obtained from ground-level sensors. But, such sensors' spatial coverage is limited, and necessitate extra deployment cost. We can estimate $PM_{2.5}$ from satellite-retrieved Aerosol Optical Depth (AOD) too. However, AOD suffers from uncertainties associated with

its retrieval algorithms and has limited spatial resolution of estimated $PM_{2.5}$. Moreover, AOD cannot be retrieved in case the weather is cloudy. On the other hand, satellite images have continuous spatial coverage without any deployment cost. Accuracy of monitoring from such satellite images is hindered due to uncertainties of sensor data of relevant environmental parameters, such as, relative humidity, temperature, wind speed and wind direction. Such uncertainties are well addressed by Belief Rule Based Expert System (BRBES). Image data are well analysed by Convolutional Neural Network (CNN). Hence, we propose to integrate CNN with BRBES to monitor air quality from satellite images more accurately while handling uncertainties. We employed two-layer BRBES in our proposed model to perform deeper reasoning over CNN output across two layers. We customized CNN and fine-tuned BRBES to increase monitoring accuracy further. We also differentiated an obscure image between polluted air and cloud in our model. Valid environmental data (temperature, wind speed, and wind direction) have been included to make the monitoring performance of our proposed model stronger. Three-year observation data (satellite images, and environmental parameters) from 2014 to 2016 of Shanghai have been employed to design and validate our proposed model. The results conclude that the monitoring accuracy of our model is higher than only CNN and other conventional Machine Learning models.

– This paper addresses the second research question of this thesis by improving accuracy and robustness of BRBES while handling data uncertainties. Two-layer BRBES of this paper contributes to the reasoning of input values by two successive layers of BRBES, resulting in more in-depth rectification. To conduct this research, we obtained 3-year satellite images of Oriental Pearl Tower, Shanghai, China from January-2014 till December-2016. These 3150 images were captured by PlanetScope, which is a constellation composed by approximately 120 optical satellites operated by Planet [36]. I was the main contributor to this research study. However, my supervisors provided me with important feedback and suggestions to improve writing as well as the overall quality of the paper.

- **Paper E.** *A Semi-Supervised-Learning-Aided Explainable Belief Rule-Based Approach to Predict the Energy Consumption of Buildings.* **Sami Kabir**, Mohammad Shahadat Hossain, and Karl Andersson. MDPI Algorithms. Published on May 23, 2025.

– Accurately predicting the energy consumption of buildings is essential for utility providers, users, and facility managers to reduce energy waste and enhance operational efficiency. However, providing such prediction with Artificial Intelligence (AI) models becomes difficult due to shortage of supervised labeled data. The acquisition of such labeled data is either costly or constrained by privacy protection. To address this scarcity of balanced labeled data, semi-supervised learning leverages unlabeled data. Motivated by this, a semi-supervised learning framework is proposed by this paper to train AI model. As AI model, this paper employs

Belief Rule-Based Expert System (BRBES) because of its prediction with domain knowledge while handling data uncertainties. BRBES' parameters and structure are optimized with initial labeled data through evolutionary learning, until its accuracy reaches a predefined confidence threshold. Within the semi-supervised learning paradigm, a self-training model is adopted, where this optimized BRBES pseudo-labels the unlabeled data. These unlabeled data are generated through weak and strong augmentation techniques. Then the BRBES is re-optimized with both labeled and pseudo-labeled data, leading to a semi-supervised BRBES. Finally, the proposed framework explains its predictive output in non-technical human language to the end user, resulting in user trust. This semi-supervised explainable BRBES framework is validated with a case study of Skellefteå, Sweden. Experimental results demonstrate that the proposed framework achieves $20 \pm 0.71\%$ improvement in prediction accuracy and $29 \pm 0.67\%$ enhancement in explainability compared to the state-of-the-art semi-supervised machine learning models.

– I was the main contributor of this paper, which addresses all the three research questions of this thesis. My supervisors supervised me in the right way to address the shortage of balanced labeled data in the learning mechanism of BRBES. While generating unlabeled data through augmentation, both diversity and density of the data distribution were taken into account. Upon learning from limited labeled data and pseudo-labeled data, initial BRBES transitioned to the semi-supervised BRBES. Thus, this framework, comprising semi-supervised BRBES and explanation interface, addresses all the three research questions concerning explainability, accuracy, and labeled data shortage. As limited labeled data, we used the hourly energy consumption dataset of 62 residential apartments of Skellefteå, obtained from Skellefteå Kraft of Sweden.

- **Paper F.** *A Self-Supervised Belief Rule-Based Approach to Predict and Explain the Energy Consumption of Buildings.* **Sami Kabir**, Mohammad Shahadat Hossain, and Karl Andersson. Submitted for review.

– Predictive insight of energy consumption can enable utility providers and facility managers to reduce energy waste of buildings. However, offering such prediction with Artificial Intelligence (AI) models is impeded by the scarcity of human-annotated labeled data. This scarcity is exacerbated by the cost and privacy constraints associated with data collection. This data scarcity is addressed by self-supervised learning method. This learning adopts self-defined pseudo-labels as supervision in the pretext task, and uses this learned representation for downstream prediction. Motivated by this learning paradigm, this paper proposes a self-supervised framework to train AI model without relying on manually labeled data. As AI model, this paper employs Belief Rule-Based Expert System (BRBES) for its capacity to predict with domain knowledge and deal with data uncertainties. Initially, the proposed framework synthetically generates unlabeled data through

weak and strong augmentation. Then, in the pretext task of self-supervised learning, these unlabeled data are pseudo-labeled by the BRBES. These pseudo-labeled data are utilized to optimize the parameters and structure of BRBES through evolutionary learning. This optimization yields a self-supervised BRBES with fine-tuned representation. This representation, fine-tuned with pseudo-labeled data, is transferred to the downstream task to provide prediction. Moreover, to promote human-AI trust, this framework communicates the explanation of its predictive output in non-technical human language. The proposed self-supervised explainable BRBES framework is validated with a case study from Skellefteå, Sweden. According to the experimental results, the proposed framework attains $23 \pm 0.87\%$ more accuracy and $29 \pm 0.71\%$ more explainability relative to the state-of-the-art self-supervised machine learning models.

– This paper also addresses all the three research questions of this thesis. I, with active supervision from my supervisors, proposed a novel theoretical framework in this paper to conduct learning mechanism of BRBES without any labeled data. For this purpose, this paper integrates self-supervised learning with BRBES. After pseudo-labeling the unlabeled data in the pretext tasks, BRBES learns representation from such pseudo-labeled data. Then, we transfer this BRBES, learned in a self-supervised approach, to downstream task for prediction. Thus, this novel framework, consisting of self-supervised BRBES and explanation interface, addresses all the three research questions concerning explainability, accuracy, and labeled data shortage. To evaluate the downstream accuracy of this self-supervised framework, we used the hourly energy consumption dataset of 62 residential apartments of Skellefteå as test dataset.

We now show how the aforementioned six papers collectively contribute to the overall research objective of this thesis in a flowchart in Figure 1.2. Unique contribution of each paper is demonstrated in this flowchart. Summary of this chapter is presented in the next sub-section.

1.6 Chapter Summary

This chapter introduced the main topic of the thesis, and a detailed discussion of the research motivation. We discussed the main research challenges identified from the state-of-the-art in XAI. We defined and described three research questions answered by this thesis, focused on domain knowledge based explanation, and overcoming the shortage of labeled data to improve accuracy of BRBES. Then, we described the 7-step research methodology used in this thesis. Finally, we listed the main thesis contributions of our research, with related papers discussing how each thesis contribution answers the research questions. Next chapter highlights the background and related work pertinent to this thesis.

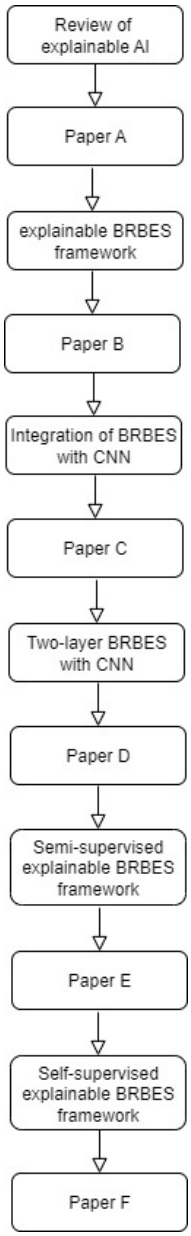

Figure 1.2: Flowchart of six papers

Chapter 2

Background and Related Work

This chapter discusses the background information and terminologies related to this thesis. Related studies to the thesis are also presented in this chapter.

2.1 Background

The following sub-section describes the background information and terminologies related to the topics covered in this thesis.

2.1.1 Explainable Artificial Intelligence

AI models are increasingly being deployed in various fields, such as, healthcare, finance, and autonomous systems. However, rising complexity of these models has triggered the demand for transparency and interpretability. Specifically, explainability of AI models is required by safety-critical application areas. AI models, particularly deep learning models, offer high predictive accuracy. However, due to opacity of the internal reasoning process, such models often function as "black-box". This opacity has led to the emergence of Explainable Artificial Intelligence (XAI) as a critical research area [7] [8]. A set of methods which provide human-understandable explanation to make the prediction of AI models transparent is called XAI [7]. Defense Advanced Research Projects Agency (DARPA) [37] defines XAI as "more explainable models, while maintaining a high level of learning performance (prediction accuracy); and enable human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners". According to Fairness, Accountability and Transparency (FAT) [38], explainability in the context of AI means "to ensure that algorithmic decisions as well as any data driving those decisions can be explained to end-users and other stakeholders in non-technical terms". In Figure 2.1, we show the role of XAI to make the decision of an AI model transparent and trustworthy to the human user. Such transparency is critical to ensure regulatory compliance in sectors directly affecting individuals, such as, finance,

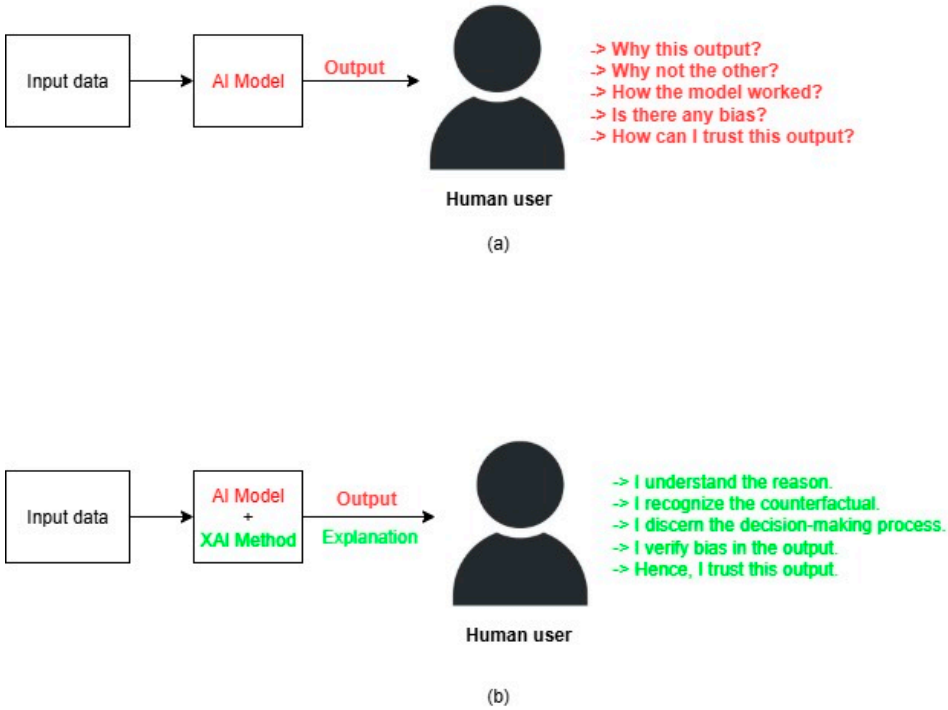

Figure 2.1: AI model predicts output for a human user: (a) without explanation, resulting in lack of trust; and (b) with explanation from an XAI method, resulting in trustworthy output.

healthcare, law, autonomous systems, and criminal justice.

To apply the XAI methods across different domains in an organized manner, a systematic taxonomy is essential. Several taxonomic frameworks have been proposed in the literature, which are generally categorized by explanation timing, explanation scope, model dependency, explanation type, audience, and interaction modality [39] [40]. Based on explanation timing, XAI techniques are classified into intrinsic (ante hoc) and post hoc methods [41]. Intrinsic (ante hoc) methods are inherently interpretable models, such as, decision trees, rule-based models, and linear regression. On the other hand, post hoc methods apply interpretability techniques after a "black-box" model's predictive output. Examples of post hoc methods include Local Interpretable Model-agnostic Explanations (LIME) [29], SHapley Additive exPlanations (SHAP) [42], Partial Dependence Plots (PDPs) [43], and counterfactual explanation. Based on the scope of explanation, XAI techniques are classified into global and local explanation [44]. Global explanation describes the overall behavior of a model across the entire input space, allowing insights

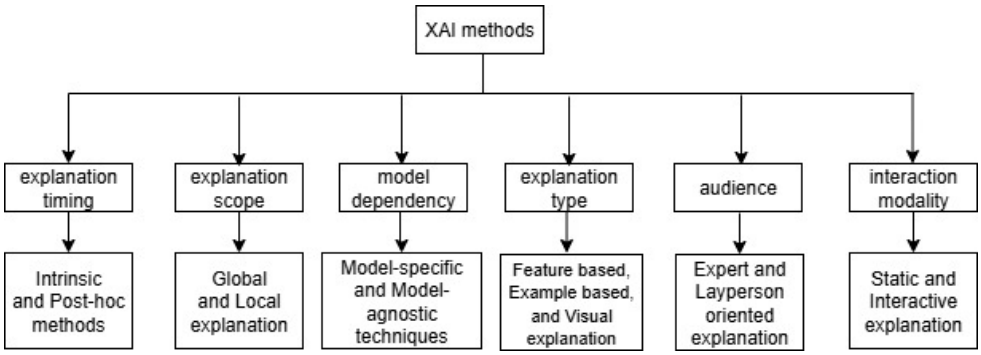

Figure 2.2: Taxonomy of XAI methods

into feature importance and model structure [40]. For example, SHAP produces global explanations by aggregating local (per-instance) shapley values of each input feature across several predictions. On the other hand, local explanation focuses on the reasoning behind individual prediction. For example, LIME explains the individual predictions of a complex model by creating a local surrogate model (e.g., linear regression or decision tree), which mimics the behavior of the complex model near the input space [29] [42]. Based on model dependency, XAI techniques are classified into model-specific and model-agnostic approaches [45]. Model-specific approaches are tailored to specific types of models. For instance, attention maps are suitable for deep learning in computer vision [46]. Model-agnostic approaches are applied universally to any AI model to extract explanations. Examples of such approaches include LIME [29], SHAP [42], and PDP [43]. Based on explanation type, XAI methods are classified into three types: feature-based, example-based, and visual explanation [47]. Feature-based explanation highlights the contribution of each input feature to the prediction [48]. Example-based explanation uses training instances to explain a prediction by highlighting its decision boundaries. Moreover, counterfactuals are generated to inform the user of preconditions to obtain an alternative outcome [49]. Visual explanation produces heat-maps over images to highlight the regions the model used to provide a prediction [50]. Based on the target audience, XAI methods are either expert-oriented or layperson-oriented [51]. Expert-oriented explanation provides detailed technical insights tailored to the developers, data scientists, or domain experts [10] [20]. On the other hand, layperson-oriented explanation is provided in non-technical language using analogies, narratives, or visuals to enhance public understanding and trust [52]. Based on interaction modality, XAI methods provide either static or interactive explanation [53]. Static explanation provides fixed reports or visualizations without user input [54]. Interactive explanation allows a user to interact with the model by changing inputs and exploring "what-if" scenarios, resulting in deeper comprehension of the model's decision-making process [55]. We show the taxonomy of XAI methods in Figure 2.2.

Despite making significant progress, the development of XAI techniques navigates through several challenges. One major challenge is the trade-off between model accuracy and explainability [56]. Highly accurate deep learning models are difficult to explain due to their complexity and opacity [57]. In contrast, simpler models, such as, decision trees or rule-based model, are easier to interpret [7]. However, such simple models may not provide as high predictive accuracy as a deep learning model on complex tasks [58]. This trade-off between accuracy and explainability is a central dilemma in XAI, which causes users to accept reduced performance in exchange for more explainable models [59]. Moreover, a good explanation depends on the context in which the AI model is deployed and on the audience receiving the explanation. For example, a financial regulator's need for explanation may be different than that of a healthcare service provider [29]. Another significant challenge is the evaluation of the explanation [60]. Unlike conventional machine learning models' metrics, such as, accuracy and F1-score, evaluating the quality of an explanation is inherently subjective and context-dependent. Various metrics have been proposed, including fidelity (how well an explanation reflects the actual decision-making process of the model) and sufficiency (whether the explanation provides enough information for a user to make an informed decision) [61]. Human factor also plays a crucial role in the intelligibility of explanations [62]. However, no universally accepted framework exists for evaluating explanations of an AI model [63]. XAI can also promote ethical AI by addressing fairness, accountability, and transparency issues. For example, by making an AI model interpretable, XAI can facilitate the identification and mitigation of discrimination and biases present in AI decisions [64]. Finally, due to regulatory requirements, there is an increasing demand for AI systems to comply with emerging laws and guidelines. For instance, the European Union (EU)'s General Data Protection Regulation (GDPR) grants "right to explanation" to an individual for an automated decision which affects the person significantly [65]. Such a legal framework creates an opportunity for XAI researchers to collaborate with policymakers and stakeholders to ensure legal compliance of AI models. To realize the objective of XAI in its precise meaning, future research has to be directed to human-centric design, interactive explanation, and standardized evaluation frameworks [7].

As XAI method, this thesis employs BRBES (**Paper B, E, and F**). Being a rule based model, BRBES falls under XAI method with intrinsic (ante hoc) explainability [10]. We describe this BRBES in the next sub-section.

2.1.2 Belief Rule Based Expert System (BRBES)

Prediction by an AI model is provided in two different approaches: data-driven approach, and knowledge-driven approach [66]. Data-driven approach includes machine learning and deep learning models, which discover actionable pattern from training data [67]. On the other hand, a knowledge-driven approach obtains a predictive output based on knowledge, which is represented by rules [68]. This approach is constituted by an expert system consisting of two components: a knowledge base and an inference engine. The knowledge base represents knowledge of the relevant domain or application area,

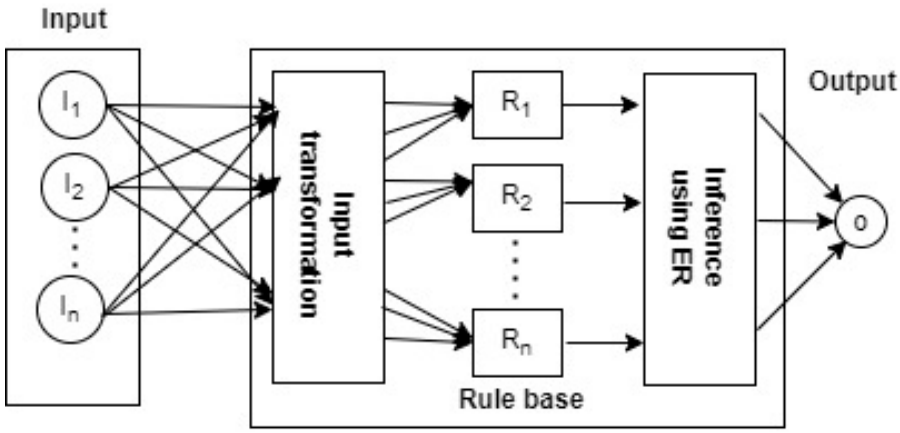

Figure 2.3: BRBES Inference Framework

and the inference engine performs reasoning over this knowledge base to infer predictive output. One prominent example of knowledge-driven approach is Belief Rule Based Expert System (BRBES) [69], which represents knowledge of the relevant domain with its rule base. It addresses various types of uncertainties, such as, incompleteness, ignorance, vagueness, imprecision, and ambiguity. It performs reasoning over the rule base with Evidential Reasoning (ER) as the inference mechanism [18]. Figure 2.3 shows a simple BRBES inference framework. This framework consists of four steps: input transformation, rule activation weight calculation, belief degree update, and rule aggregation by using ER. We now illustrate how BRBES formalizes domain knowledge with its rule base.

Rule base of BRBES extends traditional IF-THEN rule by incorporating belief structure into it. It has two parts: antecedent and consequent. The antecedent part consists of the antecedent attributes with referential values. Consequent part also has consequent attribute with belief degrees. BRBES handles various types of uncertainties with its rule weight, attribute weight, and belief degrees, as shown in Eq. (2.1).

$$R_k : \begin{cases} \text{IF } (A_1 \text{ is } V_1^k) \wedge (A_2 \text{ is } V_2^k) \wedge \ \ldots \ \wedge (A_{T_k} \text{ is } V_{T_k}^k) \\ \text{THEN } (C_1, \beta_{1k}), (C_2, \beta_{2k}), \ldots, (C_N, \beta_{Nk}) \end{cases} \tag{2.1}$$

$$\text{where } \beta_{jk} \geq 0, \sum_{j=1}^{N} \beta_{jk} \leq 1 \text{ with rule weight } \theta_k,$$

$$\text{and attribute weights } \delta_{k1}, \delta_{k2}, \ldots \delta_{kTk}, k \in 1, \ldots, L$$

Here, A_1, A_2, ... , A_{T_k} are the antecedent attributes of the k^{th} rule. V_i^k (i = 1,...,T_k, k = 1,...,L) is the referential value of the i^{th} antecedent attribute. C_j is the j^{th} referential

value of the consequent attribute. β_{jk} $(j = 1, ..., N, k = 1, ..., L)$ is the degree of belief to which the consequent referential value C_j is believed to be true. If $\sum_{j=1}^{N} \beta_{jk} = 1$ the k^{th} rule is said to be complete; otherwise, it is incomplete. T_k is the total number of antecedent attributes used in the k^{th} rule. L is the number of all belief rules in the BRBES. The total number of rules, L, in the rule base of BRBES can be calculated using the Eq. (2.2).

$$L = \prod_{i=1}^{T} J_i \tag{2.2}$$

where:

- T is the total number of antecedent attributes,
- J_i is the total number of referential (linguistic) values of the i-th antecedent attribute.

An example of a belief rule of BRBES from the energy prediction domain is given below:

IF Floor Area is Medium AND Daylight is Low AND Indoor Occupancy is High,
THEN Energy Consumption is {(High, 0.80), (Medium, 0.20), (Low, 0.0)}.

In the aforementioned rule, floor area, daylight, and indoor occupancy are the antecedent attributes, while medium, low, and high are the referential values. Energy consumption is the consequent attribute with referential values high, medium, and low. This rule is complete because the sum of degree of belief associated with each referential value of the consequent attribute is one. If the sum is less than one, then the rule will be incomplete, which may be due to incomplete information or ignorance. We now explain the four steps of the BRBES inference framework below.

In input transformation step, input data are distributed over the referential values of the antecedent attributes of a rule, which is called matching degree. Once the matching degrees are assigned, the rules become active. The combined matching degree α_k, to which the input matches the antecedent part of the k^{th} rule can be calculated by using the following equation.

$$\alpha_k = \prod_{i=1}^{T_k} (\alpha_i^k)^{\bar{\delta}_{ki}} \tag{2.3}$$

$$\text{where } \bar{\delta}_{ki} = \frac{\delta_{ki}}{\underset{i=1,...,T_k}{max}\{\delta_{ki}\}} \text{ so that } 0 \leq \bar{\delta}_{ki} \leq 1$$

Here, T_k refers to the total number of antecedent attributes in the k^{th} rule. The activation weight, w_k of the k^{th} rule is calculated with the following equation:

$$w_k = \frac{\theta_k \alpha_k}{\sum_{i=1}^{L} (\theta_i \alpha_i)} \tag{2.4}$$

Here, θ_k is the rule weight and α_k is the combined matching degree of the k^{th} rule.

Based on the matching degree of the antecedent part with input values, different rules have different activation weights. In case the matching degree is 0 (zero), the activation weight will be zero as well. If input data of any of the antecedent attributes are unavailable because of uncertainty due to ignorance, the belief degrees of each rule in the rule base will be updated. In the belief degree update step, the initial belief degree of each rule is updated using the following equation.

$$\beta_{jk} = \bar{\beta}_{jk} \frac{\sum_{t=1}^{T_k} (\lambda(t,k) \sum_{i=1}^{I_t} (\alpha_{ti}))}{\sum_{t=1}^{T_k} \lambda(t,k)} \tag{2.5}$$

$$\text{where } \lambda(t,k) = \begin{cases} 1 & \text{if the } t^{th} \text{ attribute is used in} \\ & \text{defining rule } R_k (k = 1, ..., T_k) \\ 0 & \text{otherwise} \end{cases}$$

Here, $\bar{\beta_{ik}}$ is the original belief degree, while the updated belief degree is β_{ik} in the k^{th} rule. α_{ti} represents the matching degree of input value with an antecedent attribute.

To calculate the final aggregated belief degree of each referential value of the consequent attribute, all the rules of the rule base are aggregated by using either analytical or recursive ER algorithm [70]. Analytical ER is computationally more efficient than the recursive ER [71]. Hence, we apply analytical ER in this thesis, which is calculated using the following equation [72].

$$\beta_j = \frac{\mu \times [X - \prod_{k=1}^{L} (1 - \omega_k \sum_{j=1}^{N} \beta_{jk})]}{1 - \mu \times [\prod_{k=1}^{L} 1 - \omega_k]} \tag{2.6}$$

$$\text{where } \mu = \left[\sum_{j=1}^{N} \prod_{k=1}^{L} (\omega_k \beta_{jk} + 1 - \omega_k \sum_{j=1}^{N} \beta_{jk}) - (N-1) \times \prod_{k=1}^{L} (1 - \omega_k \sum_{j=1}^{N} \beta_{jk}) \right]^{-1}$$

$$X = \prod_{k=1}^{L} (\omega_k \beta_{jk} + 1 - \omega_k \sum_{j=1}^{N} \beta_{jk})$$

Here, ω_k is the activation weight of k^{th} rule, and β_j is the belief degree of each of the referential values of the consequent attribute. The belief degrees of all the referential values of the consequent attribute can be converted into one single numerical crisp value, as shown below.

$$z_i = \sum_{j=1}^{N} u(O_j)\beta_j \tag{2.7}$$

where z_i is the numerical crisp value, and $u(O_j)$ is the utility score of the j^{th} referential value. We now show how BRBES can be made computationally effective by reducing the size of its rule base.

The rule base of BRBES, as shown in Eq. (2.1), is of conjunctive type. Such conjunctive rule base takes every possible combination of referential values of the antecedent attributes [73]. Thus, high number of referential values and/or antecedent attributes creates a large rule base under conjunctive assumption, resulting in high memory and computational requirement. To address this shortcoming, disjunctive BRBES has been proposed [74]. Disjunctive BRBES has equal number of referential values for all antecedent attributes [75]. Number of rules in the disjunctive BRBES is same as the number of referential values of its antecedent attributes. Thus, disjunctive BRBES offers small-sized rule base with lower memory capacity and computational cost. Disjunctive BRBES calculates the activation weight of its k^{th} rule, w_k, with the following equation.

$$w_k = \frac{\theta_k \sum_{i=1}^{M} \alpha_i^k}{\sum_{l=1}^{L} \theta_l \sum_{i=1}^{M} \alpha_i^l} \tag{2.8}$$

where θ_k is the initial weight of the k^{th} rule, α_i^k is the input matching degree with k^{th} rule, and L and M are the total number of rules and inputs, respectively. Disjunctive BRBES activates a rule, even if one of the input matching degrees is non-zero. On the other hand, a rule is not activated by conjunctive BRBES even if one input matching degree is zero. We now illustrate how accuracy of BRBES can be improved through learning mechanism.

Being a knowledge-driven approach, BRBES generally demonstrates lower predictive accuracy compared to machine learning and deep learning models [76]. To improve its accuracy, this study conducts learning for BRBES with supervised labeled data. As illustrated in Figure 2.4, this learning mechanism involves parallel optimization of both the parameters and the structure of BRBES. The learning parameters of BRBES comprise rule weights, antecedent attribute's weights, and consequent attribute's belief degrees [77]. These parameters are conventionally initialized by a human expert, a practice which may produce suboptimal settings for large datasets [78]. Consequently, a range of optimization techniques, such as, Particle Swarm Optimization (PSO), Differential Evolution (DE), and Sequential Quadratic Programming (SQP) have been employed to refine BRBES parameters, resulting in improved accuracy [19] [79] [80]. Although SQP-based methods can be effective for parameter tuning, these methods tend to become trapped in local optima and may therefore fail to identify the global optima within the search space [78]. By contrast, evolutionary algorithms, such as PSO, DE, Genetic Algorithms (GA), and BAT perform random searches from arbitrary locations in the search domain, which mitigates this limitation of entrapment in local optima. Owing to the optimization

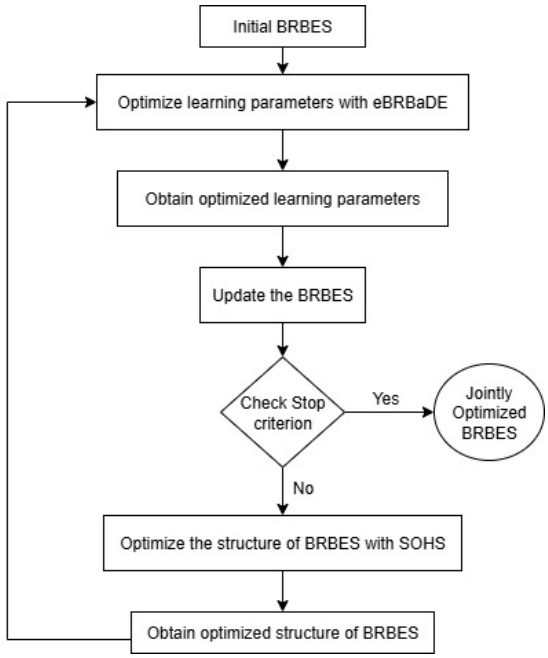

Figure 2.4: Optimization process of BRBES

strategy, DE outperforms other evolutionary algorithms to refine parameters of BRBES [81]. Two main components of DE are: exploitation, which generates new candidate solutions from the current search focus, and exploration, which expands the search domain to avoid premature convergence [82]. DE's capacity to find global optima is dependent on its two control parameters: crossover (CR) and mutation (F) [83]. Most of the DE research works have adapted the values of CR and F in response to the objective function's fitness values [84] [85]. However, these works did not consider various kinds of uncertainties inherent in the DE algorithm. To address these uncertainties of the DE algorithm, Islam et al. [78] proposed an enhanced Belief Rule-Based adaptive Differential Evolution (eBRBaDE) algorithm. The eBRBaDE algorithm leverages the BRBES' inherent uncertainty handling capacity to address the uncertainties of DE, and thereby optimize the BRBES' learning parameters. eBRBaDE contains two different BRBES components: BRBES_F and BRBES_CR, to calculate the values of F and CR respectively. In each iteration, these two BRBES components compute F and CR values based on population dynamics and objective function's values. This computation guides the search space toward an ideal balance between exploration and exploitation. eBRBaDE then generates a new population through DE mutation, crossover, and selection using the updated F and CR values. Upon completion of this process, the individual of the new population with the highest fitness value is chosen as the best option. Thus, by achieving a balance

between exploitation and exploration through uncertainty handling mechanism, eBR-BaDE provides greater accuracy than DE to find ideal values of the learning parameters of BRBES [78]. Once the parameters of BRBES is optimized with these ideal values, stop criterion of the flowchart in Figure 2.4 is examined. As this criterion is false during the first iteration, the procedure advances to the structure optimization stage. This stage finds the optimum number of referential values for each antecedent attribute of BRBES. To determine this optimum number, Structure Optimization based on Heuristic Strategy (SOHS) algorithm is employed [19]. The learning parameters associated with the SOHS-optimized structure are subsequently optimized with ideal values. Until the structure does not change further, this parameter-structure optimization loop continues. In this way, the parameters and structure of BRBES are jointly optimized by the JOPS algorithm [19]. Flowchart of this joint optimization is presented in Figure 2.4.

To identify the most optimum parameter-structure combination of BRBES, JOPS uses generalization error, rather than training error [19]. A BRBES with low training error on training dataset may perform poorly on testing dataset due to overfitting [19]. On the contrary, generalization error quantifies a predictive model's generalization capability to deliver stable results on data not encountered during training [86] [87] [88]. Moreover, according to the Hoeffding inequality theorem of probability theory [89], generalization error is a more useful criterion than training error for identifying the ideal parameters and structure of BRBES [19]. This Hoeffding inequality theorem provides an upper bound on the likelihood that the sum of random variables will differ from its expected value. Consequently, BRBES, with the minimum generalization error, is selected as the most optimal BRBES by JOPS. The generalization error, ge, is shown in the following equation.

$$ge = \frac{\sum_{i=1}^{T} \left| f(x_i) - y_i \right|}{T} + \sqrt{\frac{(d \ln 2 - \ln r)(u(D_N) - u(D_1))^2}{2T}} \tag{2.9}$$

Here, the predictive output of the BRBES for the i^{th} input data is represented by $f(x_i)$, i^{th} input data's actual output is y_i, the total number of independent input–output data pairs (x_i, y_i) is quantified by T, the total number of referential values of all antecedent attributes is symbolized by T, r is the probability value within the interval (0, 1], and the interval of the predictive output f(x) is within $[u(D_1), u(D_N)]$. We utilize labeled training dataset to optimize BRBES with JOPS. With this optimization, the initial BRBES transitions to a supervised BRBES with imrpoved accuracy, as shown in the following equation.

$$BRBES_{supervised} = JOPS(BRBES_{initial}(il)) \tag{2.10}$$

Here, *il* defines the initial labeled training data. We now illustrate how this evolutionary learning of BRBES is more explainable than the training of deep learning model.

In a traditional deep learning model, neural networks are trained with backpropagation, which calculates the gradient of a loss function with respect to the network's weights [90]. It determines the contribution of each weight to the loss. Backpropagation is used with an optimization algorithm, such as, Stochastic Gradient Descent (SGD) or

a variant of SGD to minimize the error. On the other hand, BRBES is trained with evolutionary algorithm. Main difference between the training of deep learning and evolutionary algorithm is that the evolutionary algorithm does not need backpropagation to calculate the gradient. Instead, it uses the principles of natural selection, reproduction, and genetic mutation to solve a given task [91]. In evolutionary algorithm, environment produces data. These data are used by population to make predictions. These predictive values are compared with actual output values to calculate the rewards, which are stored for each member of the population. Then, the environment ranks each member of the population according to the rewards to perform selection. Members which receive higher rewards have higher fitness. Such members are more likely to have reproductive success as well. Members, selected to reproduce, create children to form the next generation of estimators. Lastly, each child undergoes random mutation. Once mutated, each child makes prediction against the data generated by the environment. This whole process is repeated for a specific number of generations. The environment allows only the members with high fitness values to reproduce. Thus, random mutation of the new children eventually converges into an optimal solution. The fitness function of evolutionary algorithm is similar to the loss function in traditional deep learning. Evolutionary algorithm optimizes through maximization of the value of fitness function. On the other hand, traditional deep learning optimizes through minimization of the value of loss function. Hence, fitness function is a negated version of the loss function. Evolutionary approach makes the training process of BRBES more transparent compared with the training of deep learning model. Therefore, it can be argued that BRBES trains itself in a more explainable way than deep learning. Such explainable training process makes BRBES a suitable XAI method as well. We now explain the multi-layer BRBES below.

A multi-layer BRBES consists of multiple BRBESs organized hierarchically. As an example, we show a multi-layer BRBES framework to predict phenomena in Figure 2.5. This figure presents a collection of BRBESs across three different layers, which are organized hierarchically in a tree based structure. In Figure 2.5, the parent node X1 with its child nodes X2, X3, and X4 constitute BRBES X1. These child nodes are the antecedent attributes and X1 is the consequent attribute of the BRBES X1. This is the top layer or "Layer 0" of the multi-layer BRBES. BRBES X2, BRBES X3, and BRBES X4 in Figure 2.5 create the "Layer 1" of this multi-layer BRBES framework. The outputs of the BRBESs at "Layer 1" are the inputs for the BRBES X1. Similarly, the child nodes of the BRBESs at "Layer 1" get input from BRBESs of "Layer 2". Generally, the BRBESs are grouped based on relevance, and data flow in a multi-layer framework.

To provide domain knowledge based accurate prediction under uncertainty, this thesis employs BRBES (**Paper B to Paper F**). To perform reasoning across multiple layers, this thesis employs two-layer BRBES (**Paper D**). Moreover, to make BRBES more robust by dealing with multi-modal input data, this thesis integrates deep learning model with BRBES (**Paper C, and D**). A brief description of this deep learning is presented in the next sub-section.

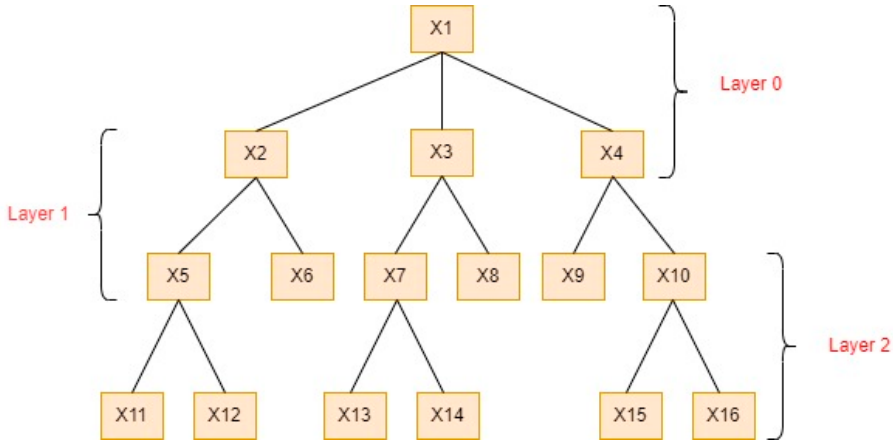

Figure 2.5: A multi-layer BRBES framework to predict phenomena

2.1.3 Deep Learning

AI exhibits machine intelligence by imitating human learning and reasoning. Alan Turing explained how a machine could imitate a human brain with "The Turing Test" in 1950 [92]. AI consists of several specific research sub-fields. Machine learning is a subset of AI, which applies statistical model to learn hidden representation of data. Deep learning is a subset of machine learning, which is based on neural network architecture with multiple hidden layers. The concept of deep learning first came to light in 2006 [93]. Initially, it was known as hierarchical learning [94].

Deep learning is mainly inspired by the interconnection among the neurons of the brain, which has two key parameters: nonlinear processing in multiple layers, and learning in supervised or unsupervised way [95]. It is a neural network, where each layer consists of several neurons. Each neuron takes input and produces output through an activation function. This activation function deals with the non-linear relationship between input and output. Neural network has an input layer, output layer, and multiple hidden layers in between. Each hidden layer learns a new feature from input data, such as, curves/edges in image recognition. Nonlinear processing across multiple layers continues with one layer taking output of the previous layer as input. This multiple layering signifies the term "deep" of deep learning. Multiple hidden layers, as opposed to a single layer, create this depth of deep learning. Deep learning models learn features of input data more deeply through these hidden layers, as shown in Figure 2.6. Weight of the connection between every two layers determines the hierarchy among these layers. This connection weight depends on the importance of the concerned input feature. Moreover, a deep learning model learns in a supervised way on labeled dataset, and unsupervised way on unlabeled dataset. Supervised learning is of two types: classification, and regression [96]. Unsupervised learning has also two types: clustering, and association [97].

Two major challenges of deep learning are overfitting and high computation time.

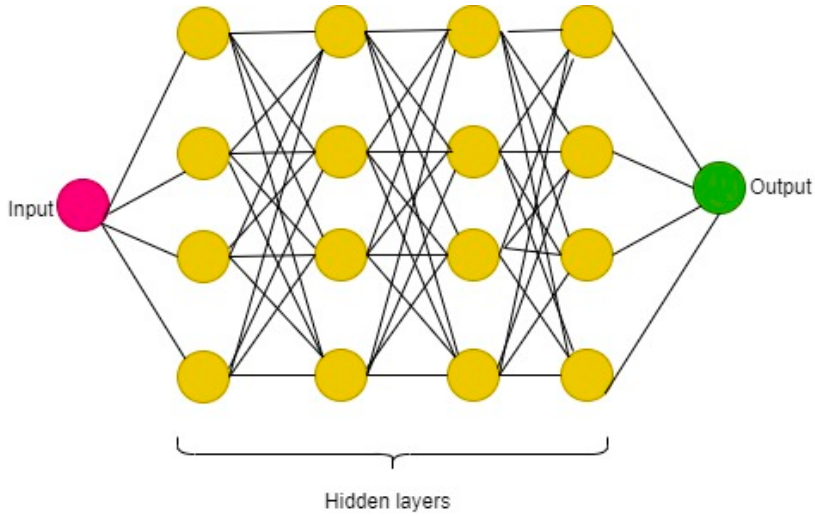

Figure 2.6: Deep Learning neural network with multiple hidden layers

A deep learning dataset is split into two parts: 80% training data, and 20% testing data [98]. Extra hidden layers of deep learning may cause overfitting with training data. To address this overfitting, various regularization methods, such as, Ivakhnenko's unit pruning, weight decay, and sparsity are applied. Moreover, some of the hidden layers are randomly removed during training phase by dropout regularization. On the other hand, number of hidden layers, number of neurons per layer, activation function calculation, and learning rate require high computation cost and time. This computation time can be reduced by various techniques, such as, batching, parallel processing, multi-core architectures, cloud computing, and high-bandwidth communication network. We now focus on the integration of deep learning with BRBES to handle multi-modal input data.

In this thesis, by dealing with multi-modal input data, BRBES has been made more robust. For handing this multi-modality, BRBES has been integrated with deep learning model [22] [23]. As deep learning model, 2-Dimensional CNN has been employed to perform analytics on visual images. Convolutional network structure is similar to the structure of animal visual cortex [99]. Individual cortical neuron receives stimuli in its own receptive field, and responds accordingly. Partially overlapped receptive fields of various neurons constitute the entire visual field.

CNN consists of multiple hidden layers between input and output layer. Convolutional layers, pooling layers, fully connected layers and normalization layers are the CNN hidden layers [100]. As activation function, CNN uses convolution and pooling functions. Convolution function takes two inputs - input image matrix, and kernel/filter matrix. This kernel matrix is employed on input image to produce an output image. Multiplica-

tion of kernel matrix with input matrix creates a modified signal. A convolution function, as a dot product of input function, f and kernel function, g, is defined as

$$(f * g)(i) = \sum_{j=1}^{m} g(j).f(i - j + m/2) \tag{2.11}$$

where m is the total number of cells, i is the current cell and j refers to all the m cells successively. Multiplication is performed between each cell of the kernel matrix with a certain cell of the input image matrix. This multiplication result is the output of convolution function, which is called convolved feature map. Size of this feature map is same as size of the kernel matrix. As activation function, Rectified Linear Unit (ReLU) is applied to this feature map. ReLU sets negative pixels to zero, and performs element-wise operation for non-linearity. Convolution and ReLU layers are applied multiple times on original image to identify hidden features of the image in a matrix form. This matrix identifies key features of an image, such as, a curve or circle of an image, while ignoring irrelevant parts of that image. This matrix is called rectified feature map [101].

Rectified feature map is given as input to the pooling function, which reduces dimensionality of this feature map with sample-based discretization. Output of the pooling function is called pooled feature map. Pooling is of two types- max pooling and min pooling. Max pooling selects the maximum pixel value of the selected sub-area of an image, as shown below.

$$h_{i,j} = \max\{x_{i+k-1,j+1-l}\} \tag{2.12}$$
$$1 \leq k \leq m \text{ and } 1 \leq l \leq m$$

where (i, j) is the position of each cell in the matrix, and m represents all the cells of concerned sub-area of the image. Max pooling is useful to focus on lighter pixels of an input image with dark background. This is particularly applicable to separate sparse features [102]. Min pooling opts for the minimum value of the concerned sub-area. This pooling focuses on dark pixels of an image with light background. After the last pooling layer, CNN has a fully connected layer. This layer takes output of the last pooling layer as input. Neurons of this fully connected layer have full connections with all neurons of the previous layer. Fully connected layer learns non-linear relationship of features extracted by previous layers to perform classification/regression. We illustrate the CNN architecture in Figure 2.7.

CNN learns image features from labeled training dataset in a supervised manner. Then CNN tests its prediction accuracy against the images of testing dataset. However, CNN cannot handle uncertainties, such as, hardware defect, ignorance, camera malfunctioning, blurred images, and camera glasses full of scratch [103]. Images captured by a camera sensor can become blurred due to turbulent weather. Snow-covered or water-marked glasses of camera will cause an image to be hazy [22]. Hence, dealing with such uncertainties of image data is of paramount importance. Here, BRBES comes into play to uphold prediction accuracy by addressing such uncertainties. CNN processes image features. BRBES processes uncertainties associated with the image features. Moreover,

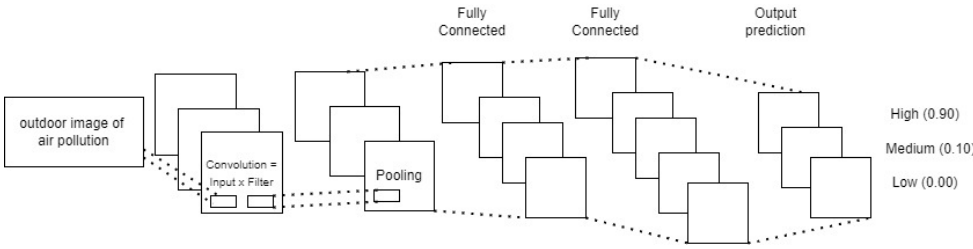

Figure 2.7: Convolutional Neural Network (CNN)

BRBES takes numerical sensor data as its another input. Thus, image data handling with CNN leads to improved accuracy and robustness of BRBES. This is the rationale of this thesis to integrate CNN with BRBES (**Paper C, and D**).

In addition to multi-modal input data handling, this thesis enhances the learning mechanism of BRBES by overcoming the shortage of labeled data. For this purpose, BRBES is integrated with semi-supervised learning (**Paper E**). We present a brief description of this semi-supervised learning method in the next sub-section.

2.1.4 Semi-Supervised Learning

To improve accuracy of BRBES, we showed its learning mechanism in sub-section 2.1.2. This evolutionary learning of BRBES requires supervised labeled data. However, in many application areas, obtaining such labeled data is challenging due to high cost, data privacy concerns [104], and manual annotation burdens [105]. A labeled dataset can be imbalanced or biased as well [10]. Such scarcity, inaccessibility, and imbalance issues hamper the proper learning of BRBES. To overcome these issues, this thesis utilizes extensive unlabeled data for learning of BRBES.

Training an AI model with a small number of labeled data points and a substantially large number of unlabeled data is called Semi-Supervised Learning (SSL) [106] [107] [108] [109]. SSL extracts knowledge from labeled data, and utilizes this extracted knowledge to assign meaningful labels to the unlabeled data. Thus, SSL utilizes the abundance of unlabeled data to improve performance of an AI model [105]. As illustrated in Figure 2.8, SSL methods are divided into five groups [110]: (a) deep generative methods, (b) consistency regularization methods, (c) graph-based methods, (d) pseudo-labeling methods, and e) hybrid methods. Examples of deep generative methods include semi-supervised Variational Autoencoders (VAEs), and Generative Adversarial Networks (GANs). Although these methods are capable of generating new training examples, attaining optimal performance simultaneously for both the generative task and downstream task is challenging. An example of consistency regularization method is Virtual Adversarial Training (VAT) [109]. As a consistency-based SSL method, a novel Semi-supervised Universal Graph Representation Learning (SUGRL) framework was proposed by reference [111].

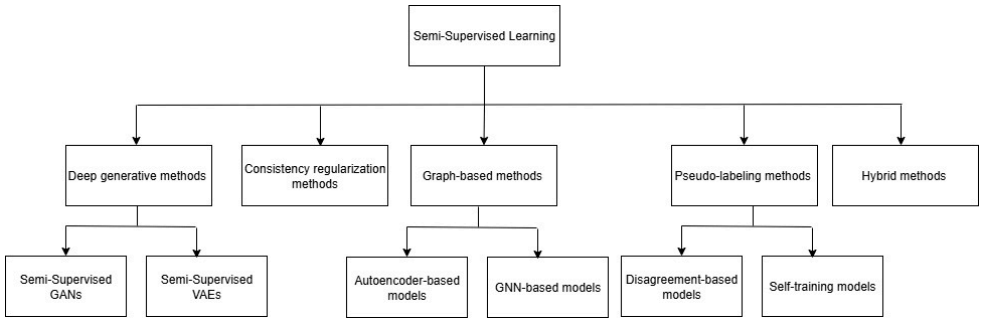

Figure 2.8: Taxonomy of semi-supervised learning methods

This framework performs semi-supervised graph classification for datasets composed of graphs from multiple, structurally heterogeneous domains. It learns domain-invariant representations by combining contrastive self-supervised learning with adversarial domain alignment. Empirical evaluation shows that this method substantially improves cross-domain generalization while relying on only a small amount of labeled data and leveraging a large amount of unlabeled graphs. Another consistency-based SSL method, Rationale-Informed Graph Neural Network (RIGNN), was proposed by the reference [112]. This method leverages task-relevant subgraphs (rationales) to enhance robustness of classification. RIGNN also improves generalization under open-world and label-sparse settings by separating discriminative information from noisy or out-of-distribution data. This study highlights the critical role of interpretability and selective representation in advancing semi-supervised graph learning. Despite the construction of consistency constraints with clear assumptions by this method, data augmentation and perturbation strategies may not always yield reliable outcome [110]. Examples of graph-based methods include Graph Neural Network (GNN), and autoencoder-based models. Nevertheless, conventional graph structures are inherently constrained to modeling pairwise relationships, which often fail to capture the higher-order dependencies of complex real-world data. To overcome this limitation, the Hypergraph-enhanced Dual Semi-supervised Graph Classification (HDSGC) framework has been proposed as an effective alternative [113]. By integrating both graph and hypergraph representations, HDSGC enables a more comprehensive modeling of intricate and high-order relationships among data instances [113]. HDSGC utilizes hyperedges to represent multi-node interactions. Such interaction enriches the contextual information available for learning. It employs a dual semi-supervised learning strategy by optimizing node classification on both graph and hypergraph views, leveraging the complementary strength of each structure. With this multi-relational representation learning using limited labeled data, this method performs graph classification with improved performance. Despite learning more information with graphs, such methods may not always build accurate relationship between training samples [110]. Examples of pseudo-labeling methods include disagreement-based models and

self-training models. As an iterative self-training strategy, SemiEvol, a semi-supervised fine-tuning approach [113], leverages both labeled and high-confidence pseudo-labeled data to progressively fine-tune Large Language Models (LLMs). SemiEvol integrates teacher-student paradigms with dynamic data selection, leading to its robust adaptation across diverse tasks under semi-supervised conditions. However, as pseudo-labels are derived from limited labeled supervision, such pseudo-labels may become noisy and potentially degrade performance [110]. Examples of hybrid methods include Interpolation Consistency Training (ICT) [114], MixMatch [115], ReMixMatch [116], and FixMatch [117]. A hybrid SSL framework was introduced through Semi-supervised Active Learning for Graph-level Classification [118], which integrates semi-supervised learning with active learning to efficiently label graph-structured data. By strategically selecting the most informative graph instances for annotation, this framework maximizes the model performance while minimizing annotation cost. By leveraging both labeled and unlabeled graphs during training, this method balances the exploration-exploitation tradeoff. This balanced tradeoff results in improved graph-level classification. Moreover, this approach integrates uncertainty estimation with structure-aware sampling strategies to attain state-of-the-art performance. Another hybrid SSL method was proposed by the reference [119], which integrates curriculum learning with pseudo-labeling in graph-based SSL to classify nodes with improved accuracy. This method iteratively pseudo-labels the unlabeled nodes based on both confidence and the complexity of the node's features. Joint consideration of confidence and complexity formulates the hybrid curriculum of this method. By employing this curriculum-guided strategy, the model gradually learns from easier to harder instances. Such gradual learning leads the model to better generalization and faster convergence. This structured learning process adapts to the evolving uncertainty of the model. Thus, this model outperforms the traditional pseudo-labeling methods [119]. Despite offering more robustness and efficiency, such hybrid SSL methods incur greater model complexity and computational overhead due to the integration of multiple learning components. Moreover, based on output, SSL is of four types [120]: semi-supervised classification, semi-supervised clustering, semi-supervised dimension reduction, and semi-supervised regression. Among these four categories, research on semi-supervised regression, aimed at predicting continuous value of a variable, remains relatively limited [121]. We now highlight the integration of SSL with BRBES below.

To overcome the shortage of labeled data in the learning mechanism, this thesis integrates SSL with BRBES [20]. As SSL, self-training method under the pseudo-labeling approach is employed. This self-training is an iterative semi-supervised learning approach, in which a supervised classifier is trained with both labeled data and data which have been pseudo-labeled in the prior iterations. The self-training procedure begins by training the supervised classifier only with the labeled data. This trained classifier is then used to predict labels for the unlabeled data. The predictive labels with the highest confidence are then appended to the training dataset. This extended training dataset, consisting of both labeled and pseudo-labeled data, are utilized to retrain the supervised classifier. This iteration typically continues until all unlabeled data are pseudo-labeled [122]. In line with this self-training method, this thesis conducts learning mechanism

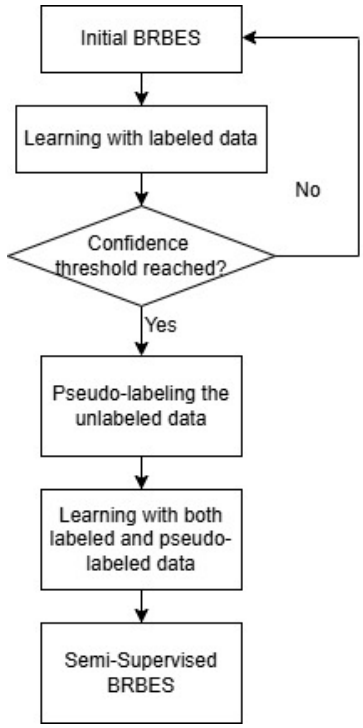

Figure 2.9: Flowchart of transition from initial to semi-supervised BRBES

of BRBES with initial labeled data until its accuracy reaches the confidence threshold. Once the threshold is reached, BRBES assigns pseudo-labels to the unlabeled data. These unlabeled data are synthetically generated through two types of augmentation: weak and strong. Weak augmentation generates integer values for relevant antecedent attributes of BRBES, which are pseudo-labeled by the BRBES (reaching the confidence threshold). To reduce class noise and attribute noise of these weakly augmented pseudo-labeled data, this thesis applies strong augmentation. As strong augmentation, this thesis generates float values for the relevant antecedent attributes of BRBES, which are then pseudo-labeled by the BRBES. These labeled as well as weakly and strongly augmented pseudo-labeled data are utilized by this thesis to conduct the learning mechanism of BRBES. Thus, by learning representation from both labeled and pseudo-labeled data, initial BRBES transitions to the semi-supervised BRBES. We show this transition process as flowchart in Figure 2.9. As application area, this thesis performs semi-supervised regression to predict continuous value of energy consumption.

This semi-supervised BRBES framework (**Paper E**), proposed by this thesis, is still dependent on limited labeled data to conduct the learning mechanism of BRBES. Hence, to perform learning of BRBES without any labeled data, this thesis integrates self-

supervised learning with BRBES (**Paper F**). This self-supervised learning method is briefly described in the next sub-section.

2.1.5 Self-Supervised Learning

Semi-supervised BRBES, as described in sub-section 2.1.4, requires limited labeled data in its learning mechanism. SSL is not completely free of labeled data [106]. Learning mechanism without any labeled data can be conducted by self-supervised learning and unsupervised learning [123] [124]. Self-supervised learning automatically generates pseudo-labels for the unlabeled data through a pre-defined pretext task [123]. These pseudo-labeled data are then utilized by the model to learn representation. Upon completion of this pre-training in the pretext task, the acquired model is transferred to the downstream task to provide prediction [123]. On the other hand, unsupervised learning discovers patterns or structure from unlabeled data through various techniques, such as, clustering, probability density, and rule discovery [124]. This learning neither uses any prior knowledge nor generates any pseudo-label [124]. Hence, to conduct learning mechanism of BRBES without any labeled data, this thesis integrates self-supervised learning with BRBES. Learning mechanism of this self-supervised BRBES is conducted with pseudo-labeled data.

A paradigm of representation learning, where a model derives supervisory signals directly from raw, unlabeled data is called self-supervised learning [125]. This learning does not need human-provided labeled data, thereby eliminating the manual annotation cost. This learning paradigm provides the supervisory signals from the raw data to the model through pretext task [126]. This pretext task is of two types: auxiliary pretext task, and contrastive learning [127]. Auxiliary pretext task learns representations from pseudo-labels, or labels created automatically based on the inherent attributes of the dataset. Examples of auxiliary pretext task include masked reconstruction (predicting masked tokens or image patches), predicting the rotation degree [128], and colorizing a grayscale image [129]. For instance, reference [130] employed auxiliary pretext task to pseudo-label the unlabeled sensor data of water distribution system in a self-supervised learning approach. Labels of the sensor devices were assigned as pseudo-labels by this study. On the other hand, contrastive learning is a discriminative model, which pulls augmented views of the same instance together, while pushing others apart [131]. This learning uses data augmentation [132], contrastive losses [133], momentum encoders [134], memory banks [135], and negative samples [135] [126]. An example of contrastive learning is a Simple Framework for Contrastive Learning of Visual Representations (SimCLR) [136], which applies strong random augmentations on the same image to create positive pairs of the image. It then utilizes these positive pairs and other negative samples to train an encoder. To perform this encoder training, SimCLR uses a contrastive loss (NT-Xent) and a projection head. Another example of contrastive learning is Momentum Contrast (MoCo) [137], which formulates contrastive learning as a dynamic dictionary lookup. It maintains a queue of encoded keys and a momentum-updated key encoder. This queue provides a large, consistent set of negative examples, thereby eliminating

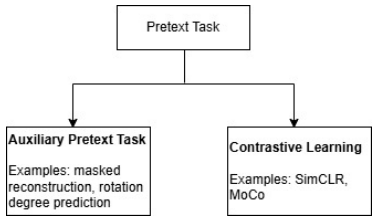

Figure 2.10: Taxonomy of pretext task of self-supervised learning

the requirement of huge synchronous batches. Pretext task enables a model to capture semantic and structural relationship from data during pretraining [125]. Therefore, the careful design of pretext task is fundamental to effective self-supervised learning. We show the taxonomy of pretext task in Figure 2.10. Once a model is pretrained on one or more pretext tasks, the learned representations are transferred to the downstream task. These representations are then evaluated or adapted for various downstream tasks, such as, classification, detection, segmentation, and regression. Thus, self-supervised learning learns generalizable, fine-tunable representations by synthesizing supervision from the unlabeled data itself. Such self-supervision has made this learning paradigm a foundational strategy to reduce manual annotation costs across many application domains [125].

As pretext task of self-supervised learning, this thesis employs auxiliary pretext task [21]. In this task, initial BRBES predicts pseudo-labels for the unlabeled data. These numerical unlabeled data are generated through weak and strong augmentation, as mentioned in sub-section 2.1.4. Then these pseudo-labeled data are utilized to conduct the learning mechanism of BRBES. Thus, by learning representation from these pseudo-labeled data, initial BRBES transitions to self-supervised BRBES. This BRBES, learned in a self-supervised approach, is then transferred to the downstream task to predict label against the input data. Downstream accuracy of this self-supervised BRBES framework, proposed by this thesis (**Paper F**), is evaluated with a labeled test dataset. We show this transition from initial BRBES to self-supervised BRBES as a flowchart in Figure 2.11. As application area of this self-supervised learning framework, this thesis performs regression in the downstream task to predict continuous value of energy consumption.

We now present the work related to this thesis, carried out by other researchers, in the next sub-section.

2.2 Related Work

This sub-section presents the related research in the field of eXplainable Artificial Intelligence (XAI), prediction with Artificial Intelligence (AI), prediction from images, semi-supervised learning, and self-supervised learning.

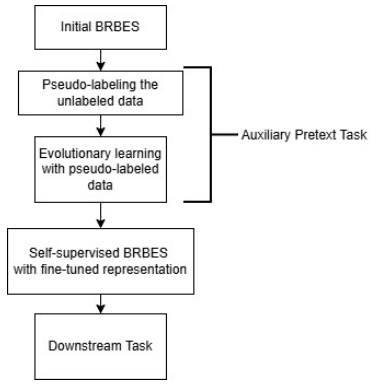

Figure 2.11: Flowchart of transition from initial to self-supervised BRBES

2.2.1 Related Work on eXplainable Artificial Intelligence (XAI)

Zhang et al. [138] proposed explanation modules to explain a Random Forest (RF)-based model. They explained feature importance with Partial Dependence Plot. They used decision tree surrogates to interpret local space, which do not take into account domain knowledge. Moreover, RF is a data-driven approach, which does not deal with data uncertainties. Tsoka et al. [139] identified the most influential input features of Artificial Neural Network (ANN) classification model with XAI tools Local Interpretable Model-agnostic Explanations (LIMEs) and Shapley Additive Explanations (SHAP). However, ANN has no domain knowledge, and does not address data uncertainties. LIMEs' local explanation and SHAP's feature importance are ad-hoc. Fan et al. [140] proposed a new LIME-based methodology to explain five different supervised models with a novel evaluation metric, trust. This metric quantitatively evaluates each prediction based on evidence type. However, a LIME-based local explanation is ad-hoc, which is dependent on training dataset, rather than domain knowledge.

Zhang et al. [141] integrated SHAP with Light Gradient Boosting Machine (LightGBM) to explain the model output with feature importance. However, LightGBM does not consider domain knowledge and data uncertainties. As a post hoc tool, SHAP gives a proxy explanation. Li et al. [142] proposed a Prediction-Explanation Network (PEN) to predict with better explainability. They employed a shared representation learning module to identify the correlation between input and output with a vector of salience. However, this PEN framework uses recurrent neural network, which does not consider domain knowledge and data uncertainties. Yu et al. [143] employed rule induction techniques to provide succinct formal explanations with background knowledge. However, this background knowledge is represented by traditional if-then rules and a boosted tree, which cannot deal with uncertainties. Chung et al. [144] applied XAI technique SHAP on a deep learning model to identify influential input parameters. However, deep learning, being a data-driven approach, does not consider relevant domain knowledge and data

uncertainties. SHAP's feature importance is also a proxy explanation.

Biessmann et al. [145] employed an Automated Machine Learning (AutoML) approach to predict the energy demand of large public buildings. Various building features and climate features were provided as input to this AutoML model. Their proposed approach showed higher prediction accuracy than the adapted and optimized Bundesinstitut fur Bau-, Stadt- und Raumforschung (BBSR) approach. However, this AutoML approach does not explain the rationale in support of the prediction, resulting in lack of trust between human and AI. Dinmohammadi et al. [146] proposed a causal inference graph, in addition to SHAP, to explain the outcome of machine learning algorithms. However, machine learning models do not incorporate domain knowledge. Hence, the feature importance, explained by the causal inference graph, is proxy. Spinnato et al. [147] proposed LASTS (Local Agnostic Subsequence-based Time Series explainer) to explain the reasons of an opaque model which classifies time series data. This explanation was made through three different representations: saliency-based explanation, instance-based explanation, and rule-based explanation. Saliency-based and instance-based explanations do not consider domain knowledge. Rules of rule-based explanation were inferred from a decision-tree surrogate, which cannot deal with data uncertainties. Guidotti et al. [148] employed stable and actionable Local Rule-based Explanation ($LORE_{sa}$) method to explain the output of a black-box model with decision tree. This decision tree made local explanation against the instance to explain. This decision tree, based on bagging-like approach, had both stability and fidelity. Ensemble decision trees were merged into a single one. Neighbor instances of the ensemble of decision trees were synthetically generated by a genetic algorithm. But, unlike BRBES, a decision tree cannot address data uncertainties.

To address the shortcomings of the aforementioned related works, this thesis sheds light on an explainable BRBES framework to provide accurate prediction and explanation based on domain knowledge, while handling data uncertainties (**Paper B**). Thus, **RQ1** is answered by this thesis. We now present the related work on prediction with various AI models in the next sub-section.

2.2.2 Related Work on Prediction with Artificial Intelligence (AI)

Theang et al. [149] proposed Dynamically pre-trained Deep Recurrent Neural Network (DRNN) to predict time-series level. Network weights of this method were continuously updated to learn representation of input data coming over time more accurately. In addition to time, they considered the position of the physical sensors to improve the prediction accuracy of DRNN. They screened out distant sensors by the elastic net method. DRNN's prediction accuracy has turned out to be higher than the autoencoder training method. However, DRNN does not handle uncertainties associated with sensor data. Li et al. [150] proposed Spatiotemporal deep learning (STDL) to predict air pollution. It extracted the spatiotemporal features of sensor data by deploying Stacked Auto Encoder (SAE). Their regression model is based on this learned representation. Their STDL model

showed higher accuracy than the Spatio-Temporal Artificial Neural Network (STANN), Auto Regression Moving Average (ARMA), and Support Vector Regression (SVR) models. However, this model also has not taken into account uncertainty regarding the sensor data. Kurt et al. [151] employed Geographic Forecasting Models using the Neural Networks (GFM_NN) method to predict the level of sulfur dioxide (SO_2), carbon monoxide (CO), and particulate matter (PM_{10}) of a Turkish district. They came up with three geographic models for prediction purpose. The first model used selected neighboring district's sensor data. The second model considered two adjoining districts. The third model calculated the distance between the triangulating districts and the target district. Their proposed geographic models showed higher accuracy than the non-geographic plain models. Among their three geographic models, third one showed better performance than the other two. However, GFM_NN has also left sensor data uncertainties unaddressed. Li, Shen, Yuan, Zhang, & Zhang [152] presented Geo-intelligent Deep Belief Network (Geoi-DBN) to predict ground-level $PM_{2.5}$ by fusing satellite and station observations. They considered geographical distance and spatiotemporal correlation in their model. $PM_{2.5}$ data of neighboring stations, nearby days of the same station, satellite retrieved Aerosol Optical Depth (AOD), Relative Humidity (RH), temperature, wind speed, surface pressure, and Normalized Difference Vegetation Index (NDVI) were taken as input features to estimate $PM_{2.5}$. However, this model does not handle sensor data uncertainties. Moreover, AOD may become erroneous due to retrieval error and cloudy weather.

To address the data uncertainties, not dealt with by the aforementioned related works, this thesis employs BRBES (**Paper B to Paper F**). Thus, **RQ2** is answered by this thesis. Related work on image based prediction is presented in the next sub-section.

2.2.3 Related Work on Prediction from Images

Li et al. [153] predicted the haze level of images with a computer vision technique. They estimated input image's transmission matrix with Dark Channel Prior (DCP) algorithm. Simultaneously, they estimated the depth map from pixels with Deep Convolutional Neural Fields (DCNF). They integrated the transmission matrix and depth map with transformation functions. Subsequently, they piled up the matrix to a single figure with pooling function and predicted haze level of the image. A combined application of transmission and depth has resulted in more accurate haze level estimation than their separate application. The accuracy of their proposed method on $PM_{2.5}$ dataset turned out to be 89.05% against 70.14% and 84.32% accuracy of only depth map and only transmission matrix, respectively. However, this method has not dealt with the uncertainties associated with captured images. Liu et al. [154] proposed an image-based approach to predict $PM_{2.5}$ of the air. They considered several image features, such as transmission, sky smoothness, image color, entropy, contrast, time, geographical location, sun, and weather condition for prediction. They developed a regression model based on these features to predict $PM_{2.5}$ level from images of Beijing, Shanghai, and Phoenix over a one-year period. Simplicity and smart phone readiness of this method can make people more aware of air pollution. However, this model also has not addressed uncertainties

associated with image data. Zhan et al. proposed a standard haze image dataset [155] containing haze images of all levels, paired with associated air quality data. Every image had Mean Opinion Score (MOS) as haze level's subjective evaluation. They also assessed the haze quality of images with a novel no-reference Image Quality Assessment (IQA) method by analyzing degradation-causing factors. IQA showed promising results on haze database, which are consistent with subjective evaluation. IQA showed better performance than Spatial and Spectral Entropies (SSEQ) and Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE). However, IQA has not taken into account image data uncertainties.

Zheng et al. [156] estimated ground-level $PM_{2.5}$ of Beijing, China from micro-satellite images by employing a combined approach of CNN and Random Forest. They extracted 128 useful features from satellite images with CNN. These 128 features, along with 4 meteorological features: temperature, RH, wind speed and sea-level pressure were given as input values to Random Forest regressor. Based on these 132 input features, this regressor predicted $PM_{2.5}$. However, Random Forest does not address uncertainties associated with sensor data of meteorological features. Albu, Precup, & Teban [157] employed Artificial Neural Network (ANN) to predict skin diseases, hepatitis B, and stroke risk. They extracted various clinical and general features of these diseases to train and test ANN. They applied ANN on Computed Tomography (CT) images to detect hepatic diseases. They also employed ANN to model finger dynamics of an amputated hand and control of a prosthetic hand. However, ANN does not deal with uncertainties associated with input features. Moreover, this method is completely data-driven, which does not take into account knowledge base from domain experts.

To address the uncertainties of image data, not handled by the aforementioned related works, this thesis integrates BRBES with CNN (**Paper C and Paper D**). With this integration, this thesis makes BRBES more robust by handling multi-modal input data. Thus, **RQ2** is answered by this thesis. We now highlight the work related to semi-supervised learning in the next sub-section.

2.2.4 Related Work on Semi-Supervised Learning

Sahoo et al. [104] employed a SSL approach, COVIDCon, to predict COVID-19 cases accurately from digital chest X-rays and Computed Tomography (CT) scans. Their proposed COVIDCon algorithm consisted of data augmentation (weak and strong), consistency regularization, and multi-contrastive learning. As weak and strong augmentation, they employed flip-shift strategy and RandAugment-based strategy respectively. However, COVIDCon algorithm's classification is based on multi-contrastive learning, with no consideration of knowledge related to COVID-19 domain. Moreover, this algorithm performs classification of images instead of regression over numerical input values. No explanation is provided by COVIDCon in support of the predictive diagnosis of COVID-19. Sahoo et al. [105] proposed a new multi-contrastive semi-supervised learning algorithm "MultiCon" to classify drug function into 12 categories by analyzing the features of two-dimensional images of drug chemical structure. This algorithm used

multi-contrastive loss with consistency regularization to predict drug function. They utilized weakly augmented versions of drug chemical structure images to pseudo-label the corresponding strongly augmented versions of the same images. However, this algorithm's multi-contrastive learning approach distinguishes image features through clustering. It does not take into account domain knowledge of drug function. Moreover, this algorithm is restricted to image classification only, rather than numerical data regression. "MultiCon" also does not explain the rationale in support of its predictive output on drug function. Li et al. [158] proposed a new semi-supervised learning methodology Mixup Contrastive Mixup (MCoM) to address the imbalanced data problem in tabular cyber security datasets. To tackle this imbalance, their proposed MCoM employed triplet mixup data augmentation on the minority (vulnerable) class to generate synthetic examples. Then they used contrastive and feature reconstruction loss to train the encoder and the decoder. They used a label propagation technique to pseudo-label the subset of the unlabeled data. For the predictor, they used Multi-Layer Perceptron (MLP). However, MLP is a data-driven approach, with no consideration of knowledge of cyber security domain. MCoM performs classification of security data, instead of regression. Moreover, security classification output is not explained by MCoM. Silva et al. [159] employed an evolutionary multi-objective algorithm, the SPEA2, for semi-supervised training of Least Squares Support Vector Machine (LSSVM) with both labeled and unlabeled data. However, LSSVM is a data driven approach, with no consideration of domain knowledge. This paper performs binary classification with LSSVM, instead of regression. Moreover, classification output is not explained by LSSVM, resulting in a lack of trust of the output to the end user. Donyavi et al. [160] proposed a model to perform classification with K-nearest neighbor (KNN). For semi-supervised learning of KNN, they proposed a synthetic, labeled data generation method called Diverse Training Dataset Generation based on a Multi-objective Optimization for Semi-Supervised Classification (DTGMO-SSC). Positions of the data, generated with DTGMO-SSC, were improved through a multi-objective evolutionary algorithm called Non-dominated Sorting Genetic Algorithm II (NSGA-II). This algorithm had two objectives: accuracy and density. The accuracy function eliminated outlier data, and the density function ensured appropriate distribution and diversity in feature space. Simultaneous consideration of accuracy and diversity was the advantage of the proposed method over the existing ones. However, KNN is a data-driven approach, which does not consider knowledge of the relevant domain. Moreover, classification output is not explained by this method. Jin et al. [161] proposed Evolutionary Optimization-based Pseudo Labeling (EOPL) for semi-supervised soft sensor development of industrial processes. They employed Gaussian Process Regression (GPR) as the base learner. Pseudo-labeling the unlabeled data was taken as an optimization problem, where pseudo-labels served as the decision variables of the Genetic Algorithm. They enlarged the labeled training dataset by combining the high-confidence pseudo-labeled data with initial labeled data. Then, the GPR model was rebuilt with this enlarged dataset. They also extended EOPL to ensemble EOPL (EnEOPL) for enhanced prediction performance. However, knowledge of industrial process domain is not taken into account by GPR. Moreover, EOPL and EnEOPL do not explain the predictive

output. The diversity of the unlabeled data in feature space is also not emphasized in this research. Gao et al. [162] proposed the Evolutionary Multi-Tasking Semi-Supervised Classification (EMT-SSC) method by combining Support Vector Machine (SVM) with a modified Z-score. This method labeled the unlabeled data with fuzzy logic and cluster assumption. However, SVM does not take into account knowledge of the relevant domain. Fuzzy logic is also outperformed by BRBES to deal with uncertainties, especially those caused by ignorance [11]. The Z-score, used by this method, may not be reliable if the training dataset does not cover all possible data instances. The proposed method performs classification, rather than regression for continuous estimation of target variable. Moreover, this method offers no explanation in support of its classification output. Cococcioni et al. [163] presented a semi-supervised-learning-aided evolutionary approach to enhance safety of workplace. They carried out a semi-supervised learning approach to formulate population for NSGA-II. Each chromosome of the population consisted of two classifiers: one classified risk perception of a worker and the other classified level of caution of the same worker. Each classifier was implemented with Multi-Layer Perceptron (MLP). This study's semi-supervised learning consisted of two stages. In the first stage, each classifier was trained independently with its own supervised training data. In the second stage, training was continued for both classifiers with an unlabeled dataset. The output of one classifier was utilized as a supervised labeled example for the other classifier. However, MLP is a data-driven approach, which does not consider the knowledge related to workplace safety. Moreover, this approach neither performs regression nor explains the classification output to the end user.

To address the shortcomings of the aforementioned related works, this thesis sheds light on a semi-supervised explainable BRBES framework (**Paper E**). This framework's BRBES component learns representation from initial labeled data. Then the framework synthetically generates unlabeled data through augmentation with proper diversity and density. Such unlabled data are then pseudo-labeled by the BRBES and utilized in its learning mechanism. Thus, by overcoming the shortage of labeled data in the learning mechanism of semi-supervised BRBES, this thesis answers **RQ3**. Moreover, other limitations of these related works, such as, domain knowledge based regression and explanation (**RQ1**), and uncertainty handling issue (**RQ2**) have been addressed by this semi-supervised explainable BRBES framework. Other researchers' work related to self-supervised learning is presented in the next sub-section.

2.2.5 Related Work on Self-Supervised Learning

Mahmoud et al. [130] proposed a two-stage detection system, consisting of both self-supervised and unsupervised algorithms, to detect Cyber-Physical (CP) attacks in water distribution systems. In the first stage, they used heuristic adaptive self-supervised algorithms using the pseudo-labeling approach to achieve near real-time decision-making. For this purpose, they used Support Vector Machine (SVM), KNN, Random Forest (RForest), Extreme Gradient Boosting (XGBoost), and Bag-of-SFA-Symbols (BOSS) on unlabeled sensor data streams. These data were collected from relevant water distribution

devices, such as, tank pressure, pump flow, pipeline flow, and pipeline status. Based on the learned pattern of water distribution and labels of these devices, they pseudo-labeled the unlabeled sensor data streams in the pretext task of self-supervised learning. Second stage validated the detected attacks using an unsupervised algorithm Isolation Forest. However, while assigning self- supervised pseudo-labels in the first stage, this work did not consider the knowledge of water distribution domain. Their proposed attack detection system performed binary classification, instead of regression over numerical input values. Moreover, no explanation is provided in support of the detected CP attack. Zhang et al. [164] proposed a novel self-supervised learning based method to detect multiple types of cyber attacks in the real-time energy market. Specifically, they employed an autoencoder-enhanced Generative Adversarial Network (GAN)-based method to examine the spatial-temporal correlations of Real-Time Locational Marginal Prices (RTLMPs) data. They employed a reconstructor to learn the spatial-temporal correlations of the normal historical RTLMP data, and reconstruct the normal RTLMPs. Through this reconstruction, they generated new labeled RTLMP data. They enhanced the reconstructor through adversarial training by discriminator. Then they constructed a detector to identify whether the RTLMPs are normal or compromised by the attackers. However, they trained reconstructor with historical RTLMPs data, instead of knowledge of the energy domain. Their proposed method classified RTLMPs data, instead of regression. Rationale in support of the classification decision is also not explained to the end-user. Almaraz-Rivera et al. [165] proposed an Intrusion Detection System (IDS) to protect Internet of Things (IoT) networks from Distributed Denial-of-Service (DDoS) attacks. Their proposed methodology generated synthetic images from flow-level IoT traffic data. Then they conducted both supervised and self-supervised learning using these images for three downstream tasks: attack detection, protocol classification, and Open Systems Interconnection (OSI) layer identification. Using supervised learning, they trained ResNet-34 architecture with images. As self-supervised learning, they employed MoCo to learn from a large quantity of negative samples. They generated unlabeled images through data augmentation incorporating random cropping, gaussian blur, and horizontal flipping. Then they labeled these unlabeled images using contrastive learning for visual representation. However, they used labeled IoT dataset to generate synthetic images, and fine-tune the model. Knowledge of the DDoS attack domain was not utilized to train their proposed model. Moreover, this study neither performs any regression nor explains the classification output. Su et al. [166] proposed a novel malware detection system by leveraging the advanced capabilities of Graph Neural Networks (GNNs) and self-supervised learning. They employed GNNs to capture malicious activity of binary files represented as graphs. To leverage unlabeled data, they adopted self-supervised learning phase, consisting of data augmentation and contrastive loss. They implemented the contrastive learning approach to learn detailed representations of benign and malicious software. With this learned representation, their proposed system labeled the unlabeled data of executable files as either malware or benign samples. However, GNN does not consider the knowledge of malware domain. The proposed system also does not explain the detection output. Huang et al. [167] proposed a new Self-Supervised Indoor

Occupancy Estimation (SSIOE) algorithm to regress the continuous value of indoor occupancy ratio (0.0-1.0) based on real-time data of four different sensors. Based on the physical interaction between the values of carbon dioxide (CO_2) concentration sensor, and power consumption sensor, they generated pseudo-labels as either high or low occupancy state. Then they trained a neural network with these sparse labels to infer the continuous occupancy ratio. They also integrated Model-Agnostic Meta-Learning (MAML) for dynamic model updating. Finally, based on the estimated indoor occupancy ratio, they built a dynamic control schedule to save energy. However, no knowledge of occupants' behavioral pattern or energy domain is considered in this research. The model's dependence on hourly sensor data requires deployment and maintenance costs of sensor devices. Moreover, rationale in support of the estimated indoor occupancy is not explained in this research. Hsu et al. [168] proposed a new deep learning model Self-supervised APPliance usage LEarning (SAPPLE) to learn when a user turns an appliance on, and location of the appliance at home. SAPPLE took input from two home sensors: a smart energy meter, and a motion sensor. They assigned labels to the unlabeled data of energy meter and motion sensor by analyzing the relationship between energy signal and location data. By learning the distribution of the residents' locations from motion sensor, the model detected the location of the appliance. They conducted the whole learning process in a self-supervised manner, without any labeled data. However, SAPPLE does not consider the knowledge of energy consumption pattern of appliance, and the residents' behavioral pattern. Moreover, no explanation is provided by SAPPLE in support of its output. Liu et al. [169] proposed a novel Transformer-based Graph Patch Informer (GPI) model to forecast renewable energy. This model has segment-wise self-attention, and graph attention networks to capture the inter-temporal dependencies of continuous signals. As training strategy, they employed self-supervised learning to enhance the representation learning. Their proposed model learned representation from the look-back window of historical data of renewable energy. With this representation, they labeled the unlabeled data of renewable energy sources. Adverse impact of missing values was also addressed by GPI. As input, time-series data of weather and meteorological variables were provided to this model. As output, the model provided wind speed forecasting, wind power forecasting, solar radiation forecasting, and photovoltaic power forecasting. However, GPI is trained with historical data, consisting of long look-back window size. Knowledge of renewable energy domain is not considered by GPI to provide forecasting. Moreover, GPI does not explain the reason behind its forecast.

To address the limitations of the aforementioned related works, this thesis proposes a self-supervised explainable BRBES framework (**Paper F**). This framework conducts the learning mechanism of BRBES without relying on any labeled data. Instead of labeled data, this framework synthetically generates unlabeled data through augmentation with proper diversity and density. It then assigns pseudo-labels to these unlabeled data with BRBES in the pretext tasks. These pseudo-labeled data are then utilized by BRBES to learn representation. Finally, this representation, fine-tuned with pseudo-labeled data, is transferred to the downstream task to perform regression. Thus, by employing self-supervised BRBES framework, this thesis eliminates the necessity of labeled data in the

learning mechanism of BRBES. With this framework, **RQ3** is answered by this thesis. Downstream accuracy of this self-supervised framework is evaluated with a labeled test dataset. Moreover, this framework addresses other shortcomings of the related works by providing domain knowledge based regression and explanation (**RQ1**), while handling data uncertainties (**RQ2**). We now summarize this chapter in the next sub-section.

2.3 Chapter Summary

In this chapter, different technologies related to the thesis topic were introduced. A brief description of XAI was presented. In addition to it, BRBES, deep learning model, semi-supervised learning, and self-supervised learning methods were introduced. Various related research studies on XAI, prediction with AI, prediction from images, semi-supervised learning, and self-supervised learning methods were presented in the later part of the chapter. Conclusions and future work will be presented next.

Chapter 3

Conclusions and Future Work

This chapter concludes the thesis and also indicates future research direction. The conclusion is presented in light of the research questions mentioned in sub-section 1.3.

3.1 Conclusions

This thesis proposed a novel Explainable Artificial Intelligence (XAI) framework with improved learning mechanisms. In light of the first research question mentioned in sub-section 1.3, an explainable Belief Rule Based Expert System (eBRBES) framework was proposed in this thesis to explain predictive output with domain knowledge. This framework represented domain knowledge with rule base of Belief Rule Based Expert System (BRBES). Thus, instead of post-hoc training dataset, this thesis explained predictive output with domain knowledge. In addition to explanation, counterfactual was provided in non-technical human language through an user interface. This counterfactual, prepared from domain knowledge, enabled an user to understand how an opposite output could have been achieved. Such explanation and counterfactual combinedly made the predictive output trustworthy to a human user. To evaluate the balance between explainability and accuracy of an Artificial Intelligence (AI) model, this thesis proposed Belief Rule Based adaptive Balance Determination (BRBaBD) algorithm. Results of BRBaBD demonstrated higher optimal balance of BRBES between explainablity and accuracy than state-of-the-art machine learning models.

The second research question concerned accuracy and robustness of BRBES, as well as data uncertainty issues. An inaccurate output of an AI model will not be trustworthy to a user, regardless of the quality of explanation. BRBES, being a knowledge-driven approach, has lower accuracy than conventional data-driven machine learning models. To improve the accuracy of BRBES, this thesis optimized both parameters and structure of BRBES through evolutionary learning. As evolutionary learning, Joint Optimization of Parameter and Structure (JOPS) algorithm was employed. To enhance robustness of BRBES, this thesis dealt with multi-modal input data. For this purpose, Convolutional Neural Network (CNN) was integrated with BRBES. This integrated model dealt

with image data and numerical data with CNN and BRBES respectively. Moreover, to perform deeper reasoning over higher number of antecedent attributes, a two-layer BRBES was proposed. Data uncertainties issue is a severe factor which hampers the prediction accuracy of any AI model. To handle such uncertainties, this thesis utilized the distributed belief structure of the rule base and belief degree update mechanism of BRBES.

The third research question was focused on addressing the shortage of labeled data in the evolutionary learning of BRBES. Labeled data are required by JOPS algorithm to conduct the learning mechanism of BRBES. However, such labeled data become scarce due to cost and privacy issues. To overcome this scarcity of labeled data, this thesis proposed semi-supervised explainable BRBES framework. As semi-supervised learning, self-training model was employed. This framework synthetically generated unlabeled data through augmentation and pseudo-labeled these unlabeled data with BRBES. These pseudo-labeled data, along with limited labeled data, were utilized by this framework to conduct the learning mechanism of BRBES. However, this framework learned representation from limited labeled data before pseudo-labeling the unlabeled data. Hence, to conduct this learning mechanism without any labeled data at all, this thesis proposed self-supervised explainable BRBES framework. This framework pseudo-labeled the augmented unlabeled data in the pretext task of self-supervised learning. Then the BRBES component of this framework learned representation from these pseudo-labeled data. This self-supervised BRBES, with fine-tuned representation, was transferred to the downstream task to perform regression. With these semi-supervised and self-supervised frameworks, this thesis improved the learning mechanism of BRBES.

Thus, in this thesis, a novel XAI framework with improved learning mechanisms was proposed. A limitation of our research is the lack of a large scale real-time deployment of an end-to-end system, consisting of BRBES and explanation interface, to predict phenomena and explain the rationale. Future research direction of this thesis is presented in the next sub-section.

3.2 Future Work

There are several directions for future research in order to advance the work presented in this thesis. The following can be considered as the future of this research.

We plan to make BRBES more autonomous by eliminating the dependence on human expert to develop its initial rule base. For this purpose, we plan to incorporate continuous and iterative learning from experience and feedback in BRBES. Such learning will enable BRBES to develop its initial rule base autonomously without any human involvement. Moreover, in the conversational interface of XAI, two-way communication between human and AI is of paramount importance. Hence, making the explanation interface of eBRBES framework bidirectional will be an interesting piece of future work. Such bidirectional interface will enable a human user to ask eBRBES framework supplementary questions and seek advice. The usability of this explanation interface also needs to be assessed by conducting a user survey. In addition to textual explanation,

visual explanation can also be incorporated to enhance the intelligibility of explanation. We plan to develop a novel metric to evaluate any XAI method. Such metric will consider explanation, counterfactual, and accuracy level of the concerned XAI method. The number of antecedent attributes, and the size of the rule base in BRBES are chosen heuristically, which may not be the global best choice [170]. The same accuracy could have been achieved even with a lower number of antecedent attributes or rules. In future, an investigation needs to be carried out to reduce the size of the rule base to make our model computationally more cost-effective, without sacrificing accuracy. Checking bias, under-representation, fairness, and privacy in the rule base is also a point which merits further investigation. Summary of this chapter is presented in the next sub-section.

3.3 Chapter Summary

This chapter concluded the thesis by highlighting the research carried out to answer the research questions. Moreover, limitations and possible future work were discussed.

References

[1] A. O. Adebayo, M. S. Chaubey, and L. P. Numbu, "Industry 4.0: The fourth industrial revolution and how it relates to the application of internet of things (iot)," *Journal of Multidisciplinary Engineering Science Studies (JMESS)*, vol. 5, no. 2, pp. 2477–2482, 2019.

[2] A. Luis Bustamante, M. A. Patricio, and J. M. Molina, "Thinger.io: An open source platform for deploying data fusion applications in iot environments," *Sensors*, vol. 19, no. 5, 2019.

[3] L. Cui, Z. Zhang, N. Gao, Z. Meng, and Z. Li, "Radio frequency identification and sensing techniques and their applications-a review of the state-of-the-art," *Sensors*, vol. 19, no. 18, 2019.

[4] Y. N. Malek et al., "On the use of iot and big data technologies for real-time monitoring and data processing," *Procedia computer science*, vol. 113, pp. 429–434, 2017.

[5] L. Chen, C. D. Nugent, and H. Wang, "A knowledge-driven approach to activity recognition in smart homes," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 6, pp. 961–974, 2012.

[6] M. Torrisi, G. Pollastri, and Q. Le, "Deep learning methods in protein structure prediction," *Computational and Structural Biotechnology Journal*, vol. 18, pp. 1301–1310, 2020.

[7] S. Kabir, M. S. Hossain, and K. Andersson, "A review of explainable artificial intelligence from the perspectives of challenges and opportunities," *Algorithms*, vol. 18, no. 9, p. 556, 2025. DOI: 10.3390/a18090556.

[8] D. Gunning and D. W. Aha, "Darpaâs explainable artificial intelligence (xai) program," *AI Magazine*, vol. 40, no. 2, pp. 44–58, 2019. DOI: 10.1609/aimag.v40i2.2850.

[9] F. K. Došilović, M. Brčić, and N. Hlupić, "Explainable artificial intelligence: A survey," in *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*, IEEE, 2018, pp. 0210–0215.

[10] S. Kabir, M. S. Hossain, and K. Andersson, "An advanced explainable belief rule-based framework to predict the energy consumption of buildings," *Energies*, vol. 17, no. 8, 2024.

[11] M. S. Hossain, S. Rahaman, R. Mustafa, and K. Andersson, "A belief rule-based expert system to assess suspicion of acute coronary syndrome (acs) under uncertainty," *Soft Computing*, vol. 22, no. 22, pp. 7571–7586, 2018.

[12] B. G. Buchanan and E. H. Shortliffe, *Rule based expert systems: the mycin experiments of the stanford heuristic programming project (the Addison-Wesley series in artificial intelligence)*. Addison-Wesley Longman Publishing Co., Inc., 1984.

[13] M. D. Herbst, E. V. Garcia, C. D. Cooke, N. F. Ezquerra, R. D. Folks, and E. G. DePuey, "Myocardial ischemia detection by expert system interpretation of thallium-201 tomograms," in *Cardiovascular Nuclear Medicine and MRI: Quantitation and Clinical Applications*, Springer, 1992, pp. 77–88.

[14] R. Ul Islam, M. S. Hossain, and K. Andersson, "A novel anomaly detection algorithm for sensor data under uncertainty," *Soft Computing*, vol. 22, no. 5, pp. 1623–1639, 2018.

[15] J. Pearl, *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan kaufmann, 1988.

[16] L. A. Zadeh, "Fuzzy logic," *Computer*, vol. 21, no. 4, pp. 83–93, 1988.

[17] M. S. Hossain, P.-O. Zander, M. S. Kamal, and L. Chowdhury, "Belief-rule-based expert systems for evaluation of e-government: A case study," *Expert Systems*, vol. 32, no. 5, pp. 563–577, 2015.

[18] J.-B. Yang, J. Liu, J. Wang, H.-S. Sii, and H.-W. Wang, "Belief rule-base inference methodology using the evidential reasoning approach-rimer," *IEEE Transactions on systems, Man, and Cybernetics-part A: Systems and Humans*, vol. 36, no. 2, pp. 266–285, 2006.

[19] L.-H. Yang, Y.-M. Wang, J. Liu, and L. Martinez, "A joint optimization method on parameter and structure for belief-rule-based systems," *Knowledge-Based Systems*, vol. 142, pp. 220–240, 2018.

[20] S. Kabir, M. S. Hossain, and K. Andersson, "A semi-supervised-learning-aided explainable belief rule-based approach to predict the energy consumption of buildings," *Algorithms*, vol. 18, no. 6, p. 305, 2025. DOI: 10.3390/a18060305.

[21] S. Kabir, M. S. Hossain, and K. Andersson, "A self-supervised belief rule-based approach to predict and explain the energy consumption of buildings."

[22] S. Kabir, R. U. Islam, M. S. Hossain, and K. Andersson, "An integrated approach of belief rule base and deep learning to predict air pollution," *Sensors*, vol. 20, no. 7, 2020.

[23] S. Kabir, R. U. Islam, M. S. Hossain, and K. Andersson, "An integrated approach of belief rule base and convolutional neural network to monitor air quality in shanghai," *Expert Systems with Applications*, vol. 206, p. 117 905, 2022.

[24] D. M. West, *The future of work: Robots, AI, and automation*. Brookings Institution Press, 2018.

[25] K. Stöger, D. Schneeberger, and A. Holzinger, "Medical artificial intelligence: The european legal perspective," *Communications of the ACM*, vol. 64, no. 11, pp. 34–36, 2021.

[26] E. P. News. "Artificial intelligence act: Meps adopt landmark law," Accessed: Apr. 17, 2024. [Online]. Available: https://www.europarl.europa.eu/news/en/press-room/20240308IPR19015/artificial-intelligence-act-meps-adopt-landmark-law.

[27] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision-making and a âright to explanationâ," *AI magazine*, vol. 38, no. 3, pp. 50–57, 2017.

[28] J. Zhu, A. Liapis, S. Risi, R. Bidarra, and G. M. Youngblood, "Explainable ai for designers: A human-centered perspective on mixed-initiative co-creation," in *2018 IEEE conference on computational intelligence and games (CIG)*, IEEE, 2018, pp. 1–8.

[29] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[30] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5188–5196.

[31] K. Främling, "Feature importance versus feature influence and what it signifies for explainable ai," in *World Conference on Explainable Artificial Intelligence*, Springer, 2023, pp. 241–259.

[32] A. Malhi and K. Främling, "An evaluation of contextual importance and utility for outcome explanation of black-box predictions for medical datasets," in *World Conference on Explainable Artificial Intelligence*, Springer, 2023, pp. 544–557.

[33] C. R. Kothari, *Research methodology: Methods and techniques*. New Age International, 2004.

[34] S.-C. Necula, F. Dumitriu, and V. Greavu-Èerban, "A systematic literature review on using natural language processing in software requirements engineering," *Electronics*, vol. 13, p. 2055, 2024. DOI: 10.3390/electronics13202055. [Online]. Available: https://doi.org/10.3390/electronics13202055.

[35] M. J. Page et al., "The prisma 2020 statement: An updated guideline for reporting systematic reviews," *International Journal of Surgery*, vol. 88, p. 105 906, 2021. DOI: 10.1016/j.ijsu.2021.105906. [Online]. Available: https://doi.org/10.1016/j.ijsu.2021.105906.

[36] Planet Team, *Planet application program interface: In space for life on earth*, https://api.planet.com, Accessed June 28, 2021, San Francisco, CA, USA, 2017.

[37] D. Gunning. "Explainable artificial intelligence (xai), defense advanced research projects agency (darpa)," Accessed: Apr. 23, 2024. [Online]. Available: https://www.darpa.mil/program/explainable-artificial-intelligence.

[38] S. Barocas, S. Friedler, M. Hardt, J. Kroll, S. Venka-Tasubramanian, and H. Wallach, "The fat-ml workshop series on fairness, accountability, and transparency in machine learning," *Accessed: Jun*, vol. 6, 2018.

[39] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (xai)," *IEEE Access*, vol. 6, pp. 52 138–52 160, 2018. DOI: 10.1109/ACCESS.2018.2870052.

[40] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Computing Surveys (CSUR)*, vol. 51, no. 5, pp. 1–42, 2018. DOI: 10.1145/3236009.

[41] C. O. Retzlaff et al., "Post-hoc vs ante-hoc explanations: Xai design guidelines for data scientists," *Cognitive Systems Research*, vol. 86, p. 101 243, 2024. DOI: 10.1016/j.cogsys.2024.101243.

[42] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2017)*, Accessed: 28 August 2025, Long Beach, CA, USA: Curran Associates, Inc., 2017, pp. 4765–4774. [Online]. Available: https://arxiv.org/abs/1705.07874.

[43] G. Szepannek and K. LÃ¼bke, "How much do we see? on the explainability of partial dependence plots for credit risk scoring," *Argumenta Oeconomica*, vol. 1, pp. 137–150, 2023. DOI: 10.15611/aoe.2023.1.06.

[44] S. Hariharan, R. R. Rejimol Robinson, R. R. Prasad, C. Thomas, and N. Balakrishnan, "Xai for intrusion detection system: Comparing explanations based on global and local scope," *Journal of Computer Virology and Hacking Techniques*, vol. 19, pp. 217–239, 2023. DOI: 10.1007/s11416-022-00437-2.

[45] Y. Wang, "A comparative analysis of model agnostic techniques for explainable artificial intelligence," *Research Reports on Computer Science*, vol. 3, pp. 25–33, 2024. DOI: 10.5281/zenodo.10654321.

[46] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy: IEEE, 2017, pp. 618–626. DOI: 10.1109/ICCV.2017.74.

[47] G. Vilone and L. Longo, "Classification of explainable artificial intelligence methods through their output formats," *Machine Learning and Knowledge Extraction*, vol. 3, pp. 615–661, 2021. DOI: 10.3390/make3030032.

[48] A. Cremades, S. Hoyas, and R. Vinuesa, "Additive-feature-attribution methods: A review on explainable artificial intelligence for fluid dynamics and heat transfer," *International Journal of Heat and Fluid Flow*, vol. 112, p. 109 662, 2025. DOI: 10.1016/j.ijheatfluidflow.2025.109662.

[49] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the gdpr," *Harvard Journal of Law & Technology*, vol. 31, pp. 841–887, 2017. [Online]. Available: https://jolt.law.harvard.edu/assets/articlePDFs/v31/WachterMittelstadtRussell.pdf.

[50] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2014. [Online]. Available: https://arxiv.org/abs/1312.6034.

[51] C. Ford and M. T. Keane, "Explaining classifications to non-experts: An xai user study of post-hoc explanations for a classifier when people lack expertise," in *International Conference on Pattern Recognition (ICPR)*, Cham, Switzerland: Springer Nature Switzerland, 2022, pp. 246–260. DOI: 10.1007/978-3-031-20225-6_18.

[52] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017. [Online]. Available: https://arxiv.org/abs/1702.08608.

[53] A. Bertrand, T. Viard, R. Belloum, J. R. Eagan, and W. Maxwell, "On selective, mutable and dialogic xai: A review of what users say about different types of interactive explanations," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*, Hamburg, Germany: ACM, 2023, pp. 1–21. DOI: 10.1145/3544548.3581418.

[54] M. Nachtigall, L. N. Q. Do, and E. Bodden, "Explaining static analysisâa perspective," in *Proceedings of the 2019 34th IEEE/ACM International Conference on Automated Software Engineering Workshop (ASEW)*, San Diego, CA, USA: IEEE, 2019, pp. 29–32. DOI: 10.1109/ASEW.2019.00012.

[55] S. Amershi, M. Chickering, S. M. Drucker, B. Lee, P. Simard, and J. Suh, "Modeltracker: Redesigning performance analysis tools for machine learning," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*, Toronto, ON, Canada: ACM, 2014, pp. 337–346. DOI: 10.1145/2556288.2557092.

[56] A. Assis, J. Dantas, and E. Andrade, "The performance-interpretability trade-off: A comparative study of machine learning models," *Journal of Reliable Intelligent Environments*, vol. 11, p. 1, 2025. DOI: 10.1007/s40860-025-00123-4.

[57] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[58] O. Serradilla, E. Zugasti, J. Rodriguez, and U. Zurutuza, "Deep learning models for predictive maintenance: A survey, comparison, challenges and prospects," *Applied Intelligence*, vol. 52, pp. 10 934–10 964, 2022. DOI: 10.1007/s10489-022-03641-5.

[59] Z. C. Lipton, "The mythos of model interpretability," *Communications of the ACM*, vol. 59, pp. 36–43, 2016. DOI: 10.1145/3233231.

[60] P. Lopes, E. Silva, C. Braga, T. Oliveira, and L. Rosado, "Xai systems evaluation: A review of human and computer-centred methods," *Applied Sciences*, vol. 12, p. 9423, 2022. DOI: 10.3390/app12199423.

[61] M. A. Kadir, A. Mosavi, and D. Sonntag, "Evaluation metrics for xai: A review, taxonomy, and practical applications," in *Proceedings of the 2023 IEEE 27th International Conference on Intelligent Engineering Systems (INES)*, Tihany, Hungary: IEEE, 2023, pp. 000 111–000 124. DOI: 10.1109/INES59327.2023.10273568.

[62] A. Abdul, J. Vermeulen, D. Wang, B. Y. Lim, and M. Kankanhalli, "Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, Montreal, QC, Canada: ACM, 2018, pp. 1–18. DOI: 10.1145/3173574.3174156. [Online]. Available: https://doi.org/10.1145/3173574.3174156.

[63] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, 2019. DOI: 10.1016/j.artint.2018.07.007. [Online]. Available: https://doi.org/10.1016/j.artint.2018.07.007.

[64] T. Calders and S. Verwer, "Three naive bayes approaches for discrimination-free classification," *Data Mining and Knowledge Discovery*, vol. 21, no. 2, pp. 277–292, 2010. DOI: 10.1007/s10618-010-0190-x. [Online]. Available: https://doi.org/10.1007/s10618-010-0190-x.

[65] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision-making and a "right to explanation"," in *Proceedings of the 2017 ICML Workshop on Human Interpretability in Machine Learning (WHI 2017)*, ser. Proceedings of Machine Learning Research, vol. 81, PMLR, 2017, pp. 1–10. [Online]. Available: https://proceedings.mlr.press/v81/goodman18a.html.

[66] L. Chen, C. D. Nugent, and H. Wang, "A knowledge-driven approach to activity recognition in smart homes," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 6, pp. 961–974, 2011. DOI: 10.1109/TKDE.2011.44. [Online]. Available: https://doi.org/10.1109/TKDE.2011.44.

[67] H. Bhavsar and A. Ganatra, "A comparative study of training algorithms for supervised machine learning," *International Journal of Soft Computing and Engineering*, vol. 2, no. 3, pp. 74–81, 2012. [Online]. Available: https://www.ijsce.org/.

[68] R. Sun, "Robust reasoning: Integrating rule-based and similarity-based reasoning," *Artificial Intelligence*, vol. 75, no. 2, pp. 241–295, 1995. DOI: 10.1016/0004-3702(94)00028-J. [Online]. Available: https://doi.org/10.1016/0004-3702(94)00028-J.

[69] J.-B. Yang, J. Liu, J. Wang, H.-S. Sii, and H.-W. Wang, "Belief rule-base inference methodology using the evidential reasoning approach (RIMER)," *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans*, vol. 36, no. 2, pp. 266–285, 2006. DOI: 10.1109/TSMCA.2005.851270. [Online]. Available: https://doi.org/10.1109/TSMCA.2005.851270.

[70] D.-L. Xu et al., "Inference and learning methodology of belief-rule-based expert system for pipeline leak detection," *Expert Systems with Applications*, vol. 32, no. 1, pp. 103–113, 2007.

[71] J.-B. Yang and P. Sen, "A general multi-level evaluation process for hybrid madm with uncertainty," *IEEE Transactions on systems, man, and cybernetics*, vol. 24, no. 10, pp. 1458–1473, 1994.

[72] Y.-M. Wang, J.-B. Yang, and D.-L. Xu, "Environmental impact assessment using the evidential reasoning approach," *European journal of operational research*, vol. 174, no. 3, pp. 1885–1913, 2006.

[73] L. Chang, X. Ma, L. Wang, and X. Ling, "Comparative analysis on the conjunctive and disjunctive assumptions for the belief rule base," in *2016 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*, IEEE, 2016, pp. 153–156.

[74] L. Chang, Z. Zhou, Y. You, L. Yang, and Z. Zhou, "Belief rule based expert system for classification problems with new rule activation and weight calculation procedures," *Information Sciences*, vol. 336, pp. 75–91, 2016.

[75] Q. Xiong, G. Chen, Z. Mao, T. Liao, and L. Chang, "Computational requirments analysis on the conjunctive and disjunctive assumptions for the belief rule base," in *2017 International Conference on Machine Learning and Cybernetics (ICMLC)*, IEEE, vol. 1, 2017, pp. 236–240.

[76] F. K. Dosilovic, M. Brcic, and N. Hlupic, "Explainable artificial intelligence: A survey," in *Proceedings of the 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, Opatija, Croatia, May 2018, pp. 0210–0215. DOI: 10.23919/MIPRO.2018.8400040.

[77] M. S. Hossain, S. Rahaman, A. L. Kor, K. Andersson, and C. Pattinson, "A belief rule based expert system for datacenter pue prediction under uncertainty," *IEEE Transactions on Sustainable Computing*, vol. 2, no. 2, pp. 140–153, 2017. DOI: 10.1109/TSUSC.2017.2664062.

[78] R. U. Islam, M. S. Hossain, and K. Andersson, "A learning mechanism for brbes using enhanced belief rule-based adaptive differential evolution," in *2020 Joint 9th International Conference on Informatics, Electronics & Vision (ICIEV) and 2020 4th International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*, IEEE, 2020, pp. 1–10.

[79] J. B. Yang, J. Liu, D. L. Xu, J. Wang, and H. Wang, "Optimization models for training belief-rule-based systems," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 37, no. 4, pp. 569–585, 2007. DOI: 10.1109/TSMCA.2007.893478.

[80] L. Chang, J. Sun, J. Jiang, and M. Li, "Parameter learning for the belief rule base system in the residual life probability prediction of metalized film capacitor," *Knowledge-Based Systems*, vol. 73, pp. 69–80, 2015. DOI: 10.1016/j.knosys.2014.09.014.

[81] L. L. Chang, Z. J. Zhou, Y. W. Chen, T. J. Liao, Y. Hu, and L. H. Yang, "Belief rule base structure and parameter joint optimization under disjunctive assumption for nonlinear complex system modeling," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 48, no. 9, pp. 1542–1554, 2018. DOI: 10.1109/TSMC.2017.2677464.

[82] C. Blum and A. Roli, "Metaheuristics in combinatorial optimization: Overview and conceptual comparison," *ACM Computing Surveys*, vol. 35, no. 3, pp. 268–308, 2003. DOI: 10.1145/937503.937505.

[83] R. D. Al-Dabbagh, F. Neri, N. Idris, and M. S. Baba, "Algorithmic design issues in adaptive differential evolution schemes: Review and taxonomy," *Swarm and Evolutionary Computation*, vol. 43, pp. 284–311, 2018. DOI: 10.1016/j.swevo.2018.03.011.

[84] J. Liu and J. Lampinen, "A fuzzy adaptive differential evolution algorithm," *Soft Computing*, vol. 9, no. 6, pp. 448–462, 2005. DOI: 10.1007/s00500-004-0363-x.

[85] M. Leon and N. Xiong, "Greedy adaptation of control parameters in differential evolution for global optimization problems," in *Proceedings of the 2015 IEEE Congress on Evolutionary Computation (CEC)*, Sendai, Japan: IEEE, 2015, pp. 385–392. DOI: 10.1109/CEC.2015.7256910.

[86] A. R. Barron, "Approximation and estimation bounds for artificial neural networks," *Machine Learning*, vol. 14, pp. 115–133, 1994. DOI: 10.1007/BF00993164.

[87] M. Seeger, "Pac-bayesian generalisation error bounds for gaussian process classification," *Journal of Machine Learning Research*, vol. 3, pp. 233–269, 2002.

[88] V. Vapnik and O. Chapelle, "Bounds on error expectation for support vector machines," *Neural Computation*, vol. 12, pp. 2013–2036, 2000. DOI: 10.1162/089976600300015042.

[89] W. Hoeffding, "Probability inequalities for sums of bounded random variables," in *The Collected Works of Wassily Hoeffding*, N. I. Fisher and P. K. Sen, Eds., New York, NY, USA: Springer, 1994, pp. 409–426. DOI: 10.1007/978-1-4612-0865-5_26.

[90] A. Shrestha and A. Mahmood, "Review of deep learning algorithms and architectures," *IEEE access*, vol. 7, pp. 53 040–53 065, 2019.

[91] R. Storn and K. Price, "Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces," *Journal of global optimization*, vol. 11, pp. 341–359, 1997.

[92] B. J. Copeland, "The turing test," *Minds and Machines*, vol. 10, no. 4, pp. 519–539, 2000.

[93] V. Rocio, A. Mosavi, and R. Ruiz, "Deep learning: A review," *Advances in Intelligent Systems and Computing*, p. 5, 2017.

[94] A. Mosavi and A. R. Varkonyi-Koczy, "Integration of machine learning and optimization for robot learning," in *Recent global research and education: technological challenges: proceedings of the 15th international conference on global research and education inter-academia 2016*, Springer, 2017, pp. 349–355.

[95] Y. Bengio et al., "Learning deep architectures for ai," *Foundations and trends® in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[96] O. Chapelle, B. Scholkopf, and A. Zien, "Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]," *IEEE Transactions on Neural Networks*, vol. 20, no. 3, pp. 542–542, 2009.

[97] P. Harrington, *Machine learning in action*. Simon and Schuster, 2012.

[98] N. M. Ball, R. J. Brunner, A. D. Myers, and D. Tcheng, "Robust machine learning applied to astronomical data sets. i. star-galaxy classification of the sloan digital sky survey dr3 using decision trees," *The Astrophysical Journal*, vol. 650, no. 1, p. 497, 2006.

[99] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological cybernetics*, vol. 36, no. 4, pp. 193–202, 1980.

[100] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[101] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, Springer, 2014, pp. 818–833.

[102] Y.-L. Boureau, J. Ponce, and Y. LeCun, "A theoretical analysis of feature pooling in visual recognition," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 111–118.

[103] Y. Li, H. Wang, B. Ding, and H. Che, "Learning from internet: Handling uncertainty in robotic environment modeling," in *Proceedings of the 9th Asia-Pacific Symposium on Internetware*, 2017, pp. 1–9.

[104] P. Sahoo, I. Roy, R. Ahlawat, S. Irtiza, and L. Khan, "Potential diagnosis of covid-19 from chest x-ray and ct findings using semi-supervised learning," *Physical and Engineering Sciences in Medicine*, vol. 45, pp. 31–42, 2021. DOI: 10.1007/s13246-021-01029-3.

[105] P. Sahoo et al., "Multicon: A semi-supervised approach for predicting drug function from chemical structure analysis," *Journal of Chemical Information and Modeling*, vol. 60, pp. 5995–6006, 2020. DOI: 10.1021/acs.jcim.0c00767.

[106] X. J. Zhu, "Semi-supervised learning literature survey," University of Wisconsin-Madison, Department of Computer Sciences, Madison, WI, USA, Tech. Rep., 2005. [Online]. Available: https://pages.cs.wisc.edu/~jerryzhu/pub/ssl_survey.pdf.

[107] O. Chapelle, B. SchÃ¶lkopf, and A. Zien, "Book review: Semi-supervised learning (chapelle, o. et al., eds.)," *IEEE Transactions on Neural Networks*, vol. 20, no. 3, p. 542, 2009. DOI: 10.1109/TNN.2009.2015974.

[108] D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Proceedings of the Workshop on Challenges in Representation Learning, ICML*, vol. 2, Atlanta, GA, USA, Jun. 2013. [Online]. Available: https://icml.cc/2013/?page_id=164.

[109] T. Miyato, S. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: A regularization method for supervised and semi-supervised learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1979–1993, 2019. DOI: 10.1109/TPAMI.2018.2858821.

[110] X. Yang, Z. Song, I. King, and Z. Xu, "A survey on deep semi-supervised learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 8, pp. 8934–8954, 2022. DOI: 10.1109/TKDE.2022.3176648.

[111] X. Luo, Y. Zhao, Y. Qin, W. Ju, and M. Zhang, "Towards semi-supervised universal graph classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 2, pp. 416–428, 2023. DOI: 10.1109/TKDE.2022.3152215.

[112] X. Luo et al., "Rignn: A rationale perspective for semi-supervised open-world graph classification," *Transactions on Machine Learning Research*, vol. 2, pp. 1–20, 2023, Available online: https://tmlr.org/. [Online]. Available: https://openreview.net/forum?id=xyz123.

[113] Z. Liu, Y. Li, Y. Wang, and Z. Wang, "Hypergraph-enhanced dual semi-supervised graph classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, Online, Feb. 2022, pp. 7603–7611. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/20595.

[114] V. Verma et al., "Interpolation consistency training for semi-supervised learning," *Neural Networks*, vol. 145, pp. 90–106, 2022. DOI: 10.1016/j.neunet.2021.10.011.

[115] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019, pp. 5049–5059. [Online]. Available: https://papers.nips.cc/paper/2019/hash/1cd138d0499a68f4bb72bee04bbec2d7-Abstract.html.

[116] D. Berthelot et al., "Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring," in *Proceedings of the International Conference on Learning Representations (ICLR)*, Addis Ababa, Ethiopia, 2020. [Online]. Available: https://openreview.net/forum?id=HklkeR4KPB.

[117] K. Sohn et al., "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," in *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS)*, Vancouver, BC, Canada, Dec. 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/06964dce9addb1c5cb5d6e3d9838f733-Abstract.html.

[118] Y. Bai, H. Zhao, Y. Zhang, Y. Liu, and L. Xie, "Semi-supervised active learning for graph-level classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, Washington, DC, USA, Feb. 2023, pp. 7520–7528. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/25428.

[119] M. Zhang, L. Song, C. Yang, and Z. Liu, "Towards effective semi-supervised node classification with hybrid curriculum pseudo-labeling," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, New Orleans, LA, USA, 2022, pp. 2175–2185. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/hash/7f89c7f6c8143dbbcfda46a31f3cb5b9-Abstract-Conference.html.

[120] J. W. Liu, Y. Liu, and X. L. Luo, "Semi-supervised learning method," Chinese, *Chinese Journal of Computers*, vol. 38, no. 8, pp. 1592–1617, 2015. DOI: 10.11897/SP.J.1016.2015.01592.

[121] S. Han and Q. H. Han, "Review of semi-supervised learning research," Chinese, *Computer Engineering and Applications*, vol. 56, no. 12, pp. 19–27, 2020.

[122] J. E. Van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Machine Learning*, vol. 109, no. 2, pp. 373–440, 2020. DOI: 10.1007/s10994-019-05855-6.

[123] V. Rani, S. T. Nabi, M. Kumar, A. Mittal, and K. Kumar, "Self-supervised learning: A succinct review," *Archives of Computational Methods in Engineering*, vol. 30, pp. 2761–2775, 2023. DOI: 10.1007/s11831-023-09884-2. [Online]. Available: https://doi.org/10.1007/s11831-023-09884-2.

[124] M. Nadif and F. Role, "Unsupervised and self-supervised deep learning approaches for biomedical text mining," *Briefings in Bioinformatics*, vol. 22, no. 2, pp. 1592–1603, 2021. DOI: 10.1093/bib/bbab016. [Online]. Available: https://doi.org/10.1093/bib/bbab016.

[125] J. Gui et al., "A survey on self-supervised learning: Algorithms, applications, and future trends," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 9052–9071, 2024. DOI: 10.1109/TPAMI.2024.3381169.

[126] S. Albelwi, "Survey on self-supervised learning: Auxiliary pretext tasks and contrastive learning methods in imaging," *Entropy*, vol. 24, no. 4, p. 551, 2022. DOI: 10.3390/e24040551. [Online]. Available: https://doi.org/10.3390/e24040551.

[127] L. Tao, X. Wang, and T. Yamasaki, "Pretext-contrastive learning: Toward good practices in self-supervised video representation learning," *arXiv preprint arXiv:2010.15464*, 2020. [Online]. Available: https://arxiv.org/abs/2010.15464.

[128] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," *arXiv preprint arXiv:1803.07728*, 2018. [Online]. Available: https://arxiv.org/abs/1803.07728.

[129] G. Larsson, M. Maire, and G. Shakhnarovich, "Colorization as a proxy task for visual understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6874–6883. DOI: 10.1109/CVPR.2017.728.

[130] H. Mahmoud, W. Wu, and M. M. Gaber, "A time-series self-supervised learning approach to detection of cyber-physical attacks in water distribution systems," *Energies*, vol. 15, no. 3, p. 914, 2022. DOI: 10.3390/en15030914.

[131] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *European Conference on Computer Vision (ECCV)*, Cham: Springer International Publishing, 2020, pp. 776–794. DOI: 10.1007/978-3-030-58621-8_45.

[132] X. Wang and G.-J. Qi, "Contrastive learning with stronger augmentations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 5, pp. 5549–5560, 2023. DOI: 10.1109/TPAMI.2022.3175346.

[133] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, H. D. III and A. Singh, Eds., ser. Proceedings of Machine Learning Research, vol. 119, PMLR, 2020, pp. 1597–1607. [Online]. Available: https://proceedings.mlr.press/v119/chen20j.html.

[134] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *arXiv preprint arXiv:2003.04297*, 2020. [Online]. Available: https://arxiv.org/abs/2003.04297.

[135] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 3733–3742. DOI: 10.1109/CVPR.2018.00393.

[136] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the 37th International Conference on Machine Learning*, H. D. III and A. Singh, Eds., ser. Proceedings of Machine Learning Research, vol. 119, PMLR, 2020, pp. 1597–1607. [Online]. Available: http://proceedings.mlr.press/v119/chen20j.html.

[137] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 9729–9738. DOI: 10.1109/CVPR42600.2020.00975. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2020/html/He_Momentum_Contrast_for_Unsupervised_Visual_Representation_Learning_CVPR_2020_paper.html.

[138] W. Zhang, F. Liu, Y. Wen, and B. Nee, "Toward explainable and interpretable building energy modelling: An explainable artificial intelligence approach," in *Proceedings of the 8th ACM international conference on systems for energy-efficient buildings, cities, and transportation*, 2021, pp. 255–258.

[139] T. Tsoka, X. Ye, Y. Chen, D. Gong, and X. Xia, "Explainable artificial intelligence for building energy performance certificate labelling classification," *Journal of Cleaner Production*, vol. 355, p. 131 626, 2022.

[140] C. Fan, F. Xiao, C. Yan, C. Liu, Z. Li, and J. Wang, "A novel methodology to explain and evaluate data-driven building energy performance models based on interpretable machine learning," *Applied Energy*, vol. 235, pp. 1551–1560, 2019.

[141] Y. Zhang, B. K. Teoh, M. Wu, J. Chen, and L. Zhang, "Data-driven estimation of building energy consumption and ghg emissions using explainable artificial intelligence," *Energy*, vol. 262, p. 125 468, 2023.

[142] S. Li, W. Liao, Y. Chen, and R. Yan, "Pen: Prediction-explanation network to forecast stock price movement with better explainability," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, 2023, pp. 5187–5194.

[143] J. Yu, A. Ignatiev, P. J. Stuckey, N. Narodytska, and J. Marques-Silva, "Eliminating the impossible, whatever remains must be true: On extracting and applying background knowledge in the context of formal explanations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, 2023, pp. 4123–4131.

[144] W. J. Chung and C. Liu, "Analysis of input parameters for deep learning-based load prediction for office buildings in different climate zones using explainable artificial intelligence," *Energy and Buildings*, vol. 276, p. 112 521, 2022.

[145] F. Biessmann, B. Kamble, and R. Streblow, "An automated machine learning approach towards energy saving estimates in public buildings," *Energies*, vol. 16, no. 19, p. 6799, 2023.

[146] F. Dinmohammadi, Y. Han, and M. Shafiee, "Predicting energy consumption in residential buildings using advanced machine learning algorithms," *Energies*, vol. 16, no. 9, p. 3748, 2023.

[147] F. Spinnato, R. Guidotti, A. Monreale, M. Nanni, D. Pedreschi, and F. Giannotti, "Understanding any time series classifier with a subsequence-based explainer," *ACM Transactions on Knowledge Discovery from Data*, vol. 18, no. 2, pp. 1–34, 2023.

[148] R. Guidotti et al., "Stable and actionable explanations of black-box models through factual and counterfactual rules," *Data Mining and Knowledge Discovery*, pp. 1–38, 2022.

[149] B. T. Ong, K. Sugiura, and K. Zettsu, "Dynamically pre-trained deep recurrent neural networks using environmental monitoring data for predicting pm 2.5," *Neural Computing and Applications*, vol. 27, pp. 1553–1566, 2016.

[150] X. Li, L. Peng, Y. Hu, J. Shao, and T. Chi, "Deep learning architecture for air quality predictions," *Environmental Science and Pollution Research*, vol. 23, pp. 22 408–22 417, 2016.

[151] A. Kurt and A. B. Oktay, "Forecasting air pollutant indicator levels with geographic models 3 days in advance using neural networks," *Expert Systems with Applications*, vol. 37, no. 12, pp. 7986–7992, 2010.

[152] T. Li, H. Shen, Q. Yuan, X. Zhang, and L. Zhang, "Estimating ground-level pm2. 5 by fusing satellite and station observations: A geo-intelligent deep learning approach," *Geophysical Research Letters*, vol. 44, no. 23, pp. 11–985, 2017.

[153] Y. Li, J. Huang, and J. Luo, "Using user generated online photos to estimate and monitor air pollution in major cities," in *Proceedings of the 7th International Conference on Internet Multimedia Computing and Service*, 2015, pp. 1–5.

[154] C. Liu, F. Tsow, Y. Zou, and N. Tao, "Particle pollution estimation based on image analysis," *PloS one*, vol. 11, no. 2, e0145955, 2016.

[155] Y. Zhan, R. Zhang, Q. Wu, and Y. Wu, "A new haze image database with detailed air quality information and a novel no-reference image quality assessment method for haze images," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2016, pp. 1095–1099.

[156] T. Zheng, M. H. Bergin, S. Hu, J. Miller, and D. E. Carlson, "Estimating ground-level pm2. 5 using micro-satellite images by a convolutional neural network and random forest approach," *Atmospheric Environment*, vol. 230, p. 117 451, 2020.

[157] A. Albu, R.-E. Precup, and T.-A. Teban, "Results and challenges of artificial neural networks used for decision-making and control in medical applications," *Facta Universitatis, Series: Mechanical Engineering*, vol. 17, no. 3, pp. 285–308, 2019.

[158] X. Li, L. Khan, M. Zamani, S. Wickramasuriya, K. W. Hamlen, and B. Thuraisingham, "Mcom: A semi-supervised method for imbalanced tabular security data," in *Proceedings of the IFIP Annual Conference on Data and Applications Security and Privacy (DBSec 2022)*, Newark, NJ, USA: Springer International Publishing, Jul. 2022, pp. 48–67. DOI: 10.1007/978-3-031-12748-6_3.

[159] C. Silva, J. S. Santos, E. F. Wanner, E. G. Carrano, and R. H. Takahashi, "Semi-supervised training of least squares support vector machine using a multiobjective evolutionary algorithm," in *Proceedings of the IEEE Congress on Evolutionary Computation (CEC 2009)*, Trondheim, Norway: IEEE, May 2009, pp. 2996–3002. DOI: 10.1109/CEC.2009.4983282.

[160] Z. Donyavi and S. Asadi, "Diverse training dataset generation based on a multi-objective optimization for semi-supervised classification," *Pattern Recognition*, vol. 108, p. 107 543, 2020. DOI: 10.1016/j.patcog.2020.107543.

[161] H. Jin, Z. Li, X. Chen, B. Qian, B. Yang, and J. Yang, "Evolutionary optimization based pseudo labeling for semi-supervised soft sensor development of industrial processes," *Chemical Engineering Science*, vol. 237, p. 116 560, 2021. DOI: 10.1016/j.ces.2021.116560.

[162] F. Gao, W. Gao, L. Huang, J. Xie, and M. Gong, "An effective knowledge transfer method based on semi-supervised learning for evolutionary optimization," *Information Sciences*, vol. 612, pp. 1127–1144, 2022. DOI: 10.1016/j.ins.2022.08.039.

[163] M. Cococcioni, B. Lazzerini, and F. Pistolesi, "A semi-supervised learning-aided evolutionary approach to occupational safety improvement," in *Proceedings of the IEEE Congress on Evolutionary Computation (CEC 2016)*, Vancouver, BC, Canada: IEEE, Jul. 2016, pp. 3695–3701. DOI: 10.1109/CEC.2016.7744227.

[164] Z. Zhang, S. Bu, Y. Zhang, and Z. Han, "Market-level integrated detection against cyber attacks in real-time market operations by self-supervised learning," *IEEE Transactions on Smart Grid*, vol. 15, no. 4, pp. 4128–4142, 2024. DOI: 10.1109/TSG.2024.3370026.

[165] J. G. Almaraz-Rivera, J. A. Cantoral-Ceballos, and J. F. Botero, "Enhancing iot network security: Unveiling the power of self-supervised learning against ddos attacks," *Sensors*, vol. 23, no. 21, p. 8701, 2023. DOI: 10.3390/s23218701.

[166] Y. Su and X. A. Wang, "A graph neural network approach with self-supervised learning for malware detection," in *Proceedings of the 2024 IEEE 7th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, vol. 7, IEEE, Sep. 2024, pp. 775–778.

[167] S. H. Huang, T. Y. Chao, B. A. Wibisono, M. P. H. Lin, and C. C. Huang, "Ssioe: Self-supervised indoor occupancy estimation for intelligent building management," *IEEE Transactions on Automation Science and Engineering*, vol. 21, no. 3, pp. 3025–3038, 2023.

[168] C. Y. Hsu, A. Zeitoun, G. H. Lee, D. Katabi, and T. Jaakkola, "Self-supervised learning of appliance usage," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

[169] J. Liu and Y. Fu, "Renewable energy forecasting: A self-supervised learning-based transformer variant," *Energy*, vol. 284, p. 128 730, 2023.

[170] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature machine intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

Department of Computer Science, Electrical and Space Engineering (SRT)
Division of Computer Science

ISSN 1402-1544
ISBN 978-91-8048-935-5 (print)
ISBN 978-91-8048-936-2 (pdf)

Luleå University of Technology 2025

LULEÅ
UNIVERSITY
OF TECHNOLOGY