Towards Human-Centered Actionable Explainable AI-enabled Systems

DISSERTATION

zur Erlangung des Grades eines Doktors rer. pol. der Fakultät III – Wirtschaftswissenschaften, Wirtschaftsinformatik und Wirtschaftsrecht der Universität Siegen

vorgelegt von

Md Shajalal, M.Eng.

Erstkorrektor: Professor Dr. Gunnar Stevens

Zweitkorrektor: Professor Dr. Alexander Boden

Dekan: Prof. Dr. Marc Hassenzahl

Datum der Disputation: March 12, 2025

Dedicated to my beloved mother

Most. Ramija Begum

for her hard work and tireless efforts to educate her children despite facing severe financial and societal challenges.

Abstract

Recently, the applications of complex artificial Intelligence (AI) models have increased exponentially in almost every sector due to the enormous advancement of computing power and the availability of high-quality annotated data for training complex machine learning (ML) models. Generally, AI models are very complex in structure, and they often need to learn thousands, even millions, of parameters in the training phase. Though the predictions are accurate, due to the complex decision-making process, the predictions from such AI models are not understandable to users. Hence, AI systems lack explainability, transparency, and trustworthiness in making the decision explainable to the users. AI models with such complexity are often referred to as *black-box* models.

To interpret the *black-box* AI models, recently, there has been a high interest in the AI research community concentrating on extracting facts and rationale to explain the reasons behind the prediction and overall models' decision-making priorities. The field that practices interpreting complex AI models and explaining the predictions to uncover the reasons behind particular predictions is known as *eXplainable Artificial Intelligence* (XAI). Improvements in interpreting ML models have been evident in this decade. However, the current explainability techniques are helpful for AI practitioners in the way that they can employ explanations to debug and eventually improve the models' performance. However, the primary objective of XAI is to help laypeople understand the predictions by providing human-centric explanations, which will eventually increase transparency and trust and lead to faster AI adoption in real-world applications. A significant gap exists in achieving human-centered explainability for AI systems due to associated challenges, including user experience variability, context sensitivity, bias and data deficiency, and actionability.

This dissertation aims to advocate the human-understandable explainability of AI-enabled systems and introduces explainable models in different real-world application scenarios. We strive to answer research questions, including i) How can we achieve high-performance ML models addressing technical challenges, including data imbalance, data inadequacy, and model bias? ii) How do explanations vary across different application contexts? iii) What underlying facts and rationale should be considered when explaining prediction for a given context? and iv) How could we achieve actionable explanations? We adopted an exploratory, experimental approach to answering these questions by conducting a wide range of experiments, introducing explainability techniques, and demonstrating explanations in three application areas: smart home, business, and natural language processing (NLP).

After carefully selecting application scenarios considering the mentioned questions, this thesis proposed multiple high-performance ML models for energy demand forecasting, occupants' thermal comfort preference modeling, product backorder prediction, multi-class patent classification, and fake review identification. Then, it introduced explainability to provide comprehensible, actionable explanations so that users and stakeholders could understand the predictions and take necessary action accordingly. The results from a wide range of experiments demonstrated high performance compared to state-of-the-art methods. They provided explanations that capture relevant facts and rationale to make users understand the proposed ML models' predictions and overall priorities. The technical and empirical evaluation of the generated explanations for explainable AI-enabled systems highlighted what information needs to be considered and how they should be represented in explanations. The broad contribution of this thesis is three-fold: i) We achieved high-performance ML models in different application areas addressing the challenges, including data inadequacy and extreme imbalance; ii)

With a wide range of experiments, this thesis gives a holistic conclusion on what facts and rationale should be employed in generating explanations for a given application context; iii) Lastly, this dissertation highlighted how we can achieve actionable explanations so that users can take necessary actions to earn more system efficiency in the given application context.

Acknowledgements

Alhamdulillah, first and foremost, I would like to express my sincere gratitude to Almighty Allah (swt.) for allowing me to finish this dissertation in good mental and physical health.

I want to thank my two excellent supervisors, Professor Dr. Alexander Boden and Professor Dr. Gunnar Stevens, for their sincere guidance and support throughout my PhD journey. Their role in providing me with the space I needed to become an independent researcher has been crucial. Their unwavering confidence in me has been a guiding light, directing me toward achieving different milestones in my PhD research and helping me stay focused in challenging situations.

I extend my heartfelt thanks to the EU Marie Skłodowska Curie Action (MSCA) program and Fraunhofer FIT for their invaluable support. The opportunity to work as an ESR on the GECKO ITN project and the financial support for my PhD research, provided by the MSCA program, has been instrumental in my academic journey. The support and freedom to conduct research, facilitated by Fraunhofer FIT, were beyond my expectations, and I am deeply grateful for the unique opportunities it has opened up for me.

I am deeply grateful to the entire GECKO consortium for their role in my academic journey. The regular PhD training events, including summer and winter schools, organized by the consortium, have been invaluable in providing a platform for learning and exchanging ideas with fellow researchers. I would like to extend my thanks to Professor Dr. Lina Stankovic and Professor Dr. Valdimir Stankovic for their supervision and warm hospitality during my research stays at the University of Strathclyde, Glasgow, UK. I also appreciate the collaboration and support of Dr. Sebastian Denef during my secondment project at OWN GmbH in Berlin, Germany. I am also thankful to all Uni Siegen Colleagues, in-

cluding Milad, Dean, Lukas, Delong, Sima, Jenny, Mahla, and Lu, for their friendly support. I would like to acknowledge Dr. Sidra Naveed, coordinator of the GECKO project, for her support in GECKO-related administrative issues.

I was fortunate to be surrounded by some fantastic colleagues in HCED at Fraunhofer FIT and IT-Security and Consumer Computing at the University of Siegen. I want to thank all the members of both research groups for making my experience exciting and fun. I am also grateful to the head of the HCED department, Dr. René Reiners, and Prof. Dr. Britta Essing for supporting me every possible way. I am also thankful to Andrea Bernerds from the HCED group for helping me with different administrative work and organizing many business trips.

I have learned a lot from my collaborators, and together, we achieved several milestones in scientific publication. My heartfelt gratitude to my collaborators, and especially I am thankful to Dr. Md. Rezaul Karim, my former colleague at Fraunhofer FIT, currently serving as a Lead Data Scientist at ALDI SÜD, for involving me in various collaborative research projects and providing me with invaluable mentoring, motivation, and guidance in both research and real-life situations. My heartiest thanks also go to other collaborators, including Chiara Tellarini from Aalborg University, Denmark, and Dr. Hari Prasanna Das from the University of California, Berkeley, USA for their support. Lastly, I want to thank Mohammad Aminul Islam, a PhD candidate at Griffith University, and my former colleague at BAU, for his support and insightful scientific discussions on various research-related issues.

I would like to acknowledge my family, my brothers and sisters, for their love and affection, which always help me and play a crucial role behind the scenes. Maa, your tireless effort alone to educate us always reminds me to do hard work and encourages me to take challenges. I acknowledge my elder brother, Md Azizul Islam, who relentlessly fulfilled his

responsibilities to the entire family after our father's passing when I was six. I want to take this opportunity to acknowledge the support of my immediate elder brother, Md Jamil Hossain, for financially supporting me in a very tough situation back in my undergraduate studies. Your support and love have been the foundation of my academic journey, and I am deeply grateful for it.

Finally, Lila, my love, how can I express the immense challenges you handled with two children, Zunayrah and Shaoib! You three were always with me during my difficult times, and your patience and accompanying me inspired me a lot. Your immeasurable love, continuous support, and the hurdles you faced encouraged and motivated me greatly throughout this PhD journey. You are the only one outside of my research area without whom this dissertation would not have been completed. Besides, I would like to acknowledge every member of the Bangladeshi community in Siegen for their unwavering support and the pleasant moments that helped me a lot during my stay in Germany. Your presence made a significant difference in my life, and I am deeply grateful for it.

Md Shajalal

Related Publications

Parts of this thesis have already been published as conference or journal papers. One paper has been published as preprint.

1. **Md Shajalal**, Alexander Boden, and Gunnar Stevens, Delong Du, Dean-Robin Kern. 2024. Explaining AI Decisions: Towards Achieving Human-Centered Explainability in Smart Home Environments. In *Proceedings of the 2nd World Conference on eXplainable Artificial Intelligence 2024 (xAI2024)*. Communications in Computer and Information Science, Springer Nature Switzerland, 418–440. https://doi.org/10.1007/978-3-031-63803-9_23 (*Reproduced with permission from Springer Nature*)

2. **Md Shajalal**, Alexander Boden, and Gunnar Stevens. 2022. Towards User-centered Explainable Energy Demand Forecasting Systems. In *Proceedings of the Thirteenth ACM International Conference on Future Energy Systems, (e-Energy '22)*. Association for Computing Machinery, New York, NY, USA, 446 – 447. https://doi.org/10.1145/3538637.3538877

3. **Md Shajalal**, Alexander Boden, Gunnar Stevens. 2024. ForecastExplainer: Explainable household energy demand forecasting by approximating shapley values using DeepLIFT. *Technological Forecasting and Social Change*, Elsevier. Volume 206, 2024, 16 pages. https://doi.org/10.1016/j.techfore.2024.123588

4. **Md Shajalal**, Milad Bohlouli, Hari Prasanna Das, Alexander Boden, and Gunnar Stevens 2022. Focus on What Matters: Improved Feature Selection Techniques for Personal Thermal Comfort Modelling, In *Proceedings of the 9th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (BuildSys '22)*. Association for Computing Machinery, New York,

NY, USA, 496 – 499. https://doi.org/10.1145/3563357.3567406

5. © 2024 IEEE. Reprinted, with permission, from **Md Shajalal**, Milad Bohlouli, Hari Prasanna Das, Alexander Boden, and Gunnar Stevens. 2024. Improved Thermal Comfort Model Leveraging Conditional Tabular GAN Focusing on Feature Selection. *IEEE Access*, IEEE vol. 12, 30039–30053. https://doi.org/10.1109/ACCESS.2024.3366453

6. **Md Shajalal**, Alexander Boden and Gunnar Stevens. 2022. Explainable Product Backorder Prediction Exploiting CNN: Introducing Explainable Models in Businesses. *Electronic Markets*, Springer-Nature 32, 2107–2122. https://doi.org/10.1007/s12525-022-00599-z (*Reproduced with permission from Springer Nature*)

7. **Md Shajalal**, Sebastian Denef, Md. Rezaul Karim, Alexander Boden and Gunnar Stevens. 2023. Unveiling the Black Box: Explainable Deep Learning Models for Patent Classification. In *Proceedings of the 2nd World Conference on eXplainable Artificial Intelligence 2023 (xAI2023)*. Communications in Computer and Information Science, Springer Nature Switzerland 457–474. https://doi.org/10.1007/978-3-031-44067-0_24 (*Reproduced with permission from Springer Nature*)

8. **Md Shajalal**, Md Atabuzzaman, Alexander Boden, Gunnar Stevens and Delong Du. 2024. What Matters in Explanations: Towards Explainable Fake Review Detection Focusing on Transformers, *Preprint in ArXiv*, 2024, 1-10. https://doi.org/10.48550/arXiv.2407.21056

Moreover, these publications contribute to the presented topic. However, they are not included as chapter of this thesis.

1. Chiara Tellarini, **Md Shajalal**, Nico Castelli, Martin Stein, Alexander Boden and Toke Haunstrup Christensen. 2024. A mixed-method approach to study the impacts of energy micro-generation combined with appliance-level feedback on everyday practices. *Energy Efficiency*, 17 (94) Springer.
   https://doi.org/10.1007/s12053-024-10276-z.

2. Md Rezaul Karim, Tanhim Islam, **Md Shajalal**, Oya Beyan, Christoph Lange, Michael Cochez, Dietrich Rebholz Schuhmann and Stefan Decker. 2023. Explainable AI for Bioinformatics: Methods, Tools, and Applications. *Briefings in Bioinformatics*, Oxford University Press 2023, 24(5), 1–22. https://doi.org/10.1093/bib/bbad236

3. Md Atabuzaman, **Md Shajalal**, Maksuda Bilkis Baby and Alexander Boden. 2023. Arabic Sentiment Analysis with Noisy Explainable Deep Learning model. In *the proceedings of the ACM Natural Language Processing and Information Retrieval* (ACM NLPIR 2023). Association for Computing Machinery, New York, NY, USA, 2023, 185–189. https://dl.acm.org/doi/10.1145/3639233.3639241

4. Md. Rezaul Karim, **Md Shajalal**, Alex Graß, Till Döhmen, Sisay Adugna Chala, Alexander Boden, Christian Beecks and Stefan Decker. 2023. Interpreting Black-box Models for High Dimensional Datasets. In *the proceedings of the 10th IEEE International Conference on Data Science and Advanced Analytics* (IEEE DSAA2023). IEEE, 2023, 1-10. https://doi.org/10.1109/DSAA60987.2023.10302562

5. Md. Rezaul Karim, Lina Comet, **Md Shajalal**, P. de Perthuis, D. Rebholz-Schuhmann, Stefan Decker. 2023. From Large Language Models (LLMs) to Knowledge Graphs for Biomarker Discovery in

Cancer. In *the proceedings of the 57th Hawaii International Conference on System Sciences* (HICSS 2023), Hawai, USA https://hdl.handle.net/10125/107055

6. Dean-Robin Kern, Gunnar Stevens, Erik Dethier, Sidra Naveed, Fatemeh Alizadeh, Du Delong, **Md Shajalal**, "Peeking Inside the Schufa Blackbox: Explaining the German Housing Scoring System", *ACM CHI 2023 Workshop on Human-Centered Explainable AI* (HCXAI), collocated with *ACM CHI Conference on Human Factors in Computing Systems 2023*, Hamburg, Germany. https://arxiv.org/abs/2311.11655

7. Lu Jin, Alexander Boden, and **Md Shajalal**. 2022. Automated Decision Making Systems in Smart Homes: A Study on User Engagement and Design. *AutomationXP22: Engaging with Automation collocated the ACM CHI Conference on Human Factors in Computing Systems 2022*, New Orleans, LA. https://ceur-ws.org/Vol-3154/paper12.pdf

Contents

List of Figures

List of Tables

Part I

Introduction & Overview

1 Introduction

1.1 Motivation

In the past decade, the availability of Artificial Intelligence (AI) applications has exponentially increased and covers almost every aspect of our lives. The applications cover almost all sectors, from entertainment to medication and diagnostics and from smart homes to autonomous vehicles [127, 278, 218]. Hence, AI applications are becoming ubiquitous these days. These significant developments have happened due to the enormous growth of the advancement in computing power and the development of high-performance, sophisticated ML algorithms, especially in deep learning (DL).

Along with that, the availability of vast amounts of high-quality data through the advancement of accurate sensing power, for example, advanced web platforms, text and image data from social media, and high-quality sensor data from connected home technology, make it possible to analyze and provide unprecedented faster and accurate predictions. Hence, AI-enabled systems are leading to breakthroughs in different areas, including smart homes, computer vision, natural language processing (NLP), and predictive analysis in business and biomedical domains.

Moreover, the widespread availability of data and the implementation of internet-connected devices have pushed the boundary of expansion of AI-enabled systems in diverse sectors, especially in healthcare [228, 307], the Internet of Things (IoT) [10], manufacturing [338], education [36, 272], and finance [9, 193, 91]. AI is revolutionizing disease diagnosis and medication recommendation [228]. In another sector, for example, in manufacturing, AI-enabled systems can monitor and streamline production systems [338], while in education, it can recommend required lessons through personalized assistive learning platforms for students [272]. The evolution of such AI systems is extending

exponentially; hence, more improved and more comfortable human lives are becoming increasingly apparent, which signals a new era of innovation and transformations in every sector.

The widespread use of the recently introduced Large Language Models (LLMs) in diverse sectors is a testament to the relentless efforts of the NLP research community [344]. With their unprecedented ability to learn the context and tone of particular users, these systems have revolutionized the field. They can generate human-like text for various NLP tasks, from text summarizing to creative writing [326]. The applications of LLMs are now typical in terms of empowered virtual assistants, content generation platforms, and even facilitating communication for people with speech impairments. LLM is not just an application of NLP research - it extends beyond language-centric applications. It is now employed in healthcare, education, and finance through data analysis, decision-making, and personalized platforms [324, 307], all thanks to the dedication of the NLP research community.

1.2 Black-Box AI and XAI

Though the emerging ML models are highly accurate and efficient in decision-making based on knowledge learned from massive fine-grained datasets, these models lack interpretability, transparency, and trust on the user end [4, 150]. The primary reason behind this is the models' high complexity to make particular decisions. ML models, especially DL, and their variants are more complex than laypeople can think of, and input values go through many complex calculations. On top of that, DL models often need to learn thousands or even millions of parameters for convergence [151]. Hence, the learning and decision-making process is like a "*black-box*" to the users.

The trained ML model is supposed to be a magic box where the user

provides inputs, and promptly, the model provides the predicted output. Therefore, the model's decision-making priorities and the underlying mechanism are not understandable and might not make users sense. Moreover, the reason behind the particular decision or prediction is unknown to the users, which might lead to mistrust. Hence, it can create issues related to the real-world adoption of such complex models.

Recently, there has been massive attention to opening the black-box models and explaining the overall models' decision-making behind the specific prediction from such model presenting the facts, priorities, and rationale [251]. The field that practices and investigates to uncover the black-box models and explain or interpret models' overall decision-making priorities and specific prediction visualizing by highlighting the facts, reasons, priorities, and rationale is refereed to *eXplainable Artificial Intelligence (XAI)*[1] [209, 251]. The terms *explainability* and *interpretability* are broadly used interchangeably in the XAI research community since both focus on uncovering the complex decision-making procedure and the facts behind their predictions.

Various explainability techniques have recently been introduced following different contexts and scenarios. These techniques can be broadly classified as *global* and *local*, where global XAI techniques focus on explaining the overall models' priorities, while local XAI explains the specific prediction [195, 244]. Based on the generic explainability, XAI techniques can be classified as model-agnostic and model-specific. Model-agnostic techniques can explain the predictions from any ML model while model-specific can only explain prediction for a specific ML model [244]. Other than that, multiple hypotheses and contexts have been considered to represent and explain the predictions, including *example-based*, *what-if* scenario-based, contrastive, and *counterfactual* explanations. So far, most widely applied explainability techniques

[1] *Explainability* and *interpretability* is used interchangeably throughout the dissertation.

include SHAP (Shapley additive explanations) [195], LIME (Local Interpretable Model-agnostic Explanation) [244], DeepLIFT (Deep Learning Important FeaTure) [293], LRP (Layer-wise Relevance Propagation) [211], Grad-Cam [271], ICE (Individual Conditional Expectation) [45], DiCE (diverse counterfactual explanations) [213], and PDP (partial dependency plot) [110]. Combining different primary XAI techniques, some notable XAI frameworks can be used as a library to apply for explaining individual ML model decisions, including Captum [166], AIX360 [24], and Anchor [246].

1.3 Towards Human-centered XAI

The significant efforts in making AI models explainable so far are for the AI practitioners and technical developers, where the intention behind explainability is to modify the model to improve the performance and debugging to find out possible errors [251, 277]. The progress in explaining complex ML models and the decision-making process is impressive from the technical perspective but needs to be closer to providing explanations understandable to general users. In other words, the explanations from state-of-the-art XAI techniques help AI practitioners improve the models' performance. However, to make explanations human-centered, much effort is needed to solve challenges related to achieving user-centered explainability. The existing literature on the efficiency of generating general user-understandable explanations concluded that the prominent XAI techniques, including SHAP and LIME and other techniques, fail to make user sense in various application scenarios [251, 38].

The forms, types, and representation of explanations varied widely across different contexts, the level of expertise of the users, and the application scenarios [251]. We can use the revised version of the famous phrase here to describe the situation with "*one XAI technique does not*

fit all" [104]. Different applications have diverse contexts and pragmatics, and most importantly, the consumers of explanations, the general users, and the stockholders of different levels of expertise require various information needs through explanations [173]. Therefore, human-understandable explainability needs to address several challenges related to human-computer interaction, user study, and technical challenges [200]. The notable technical challenges to achieving robust explainability include the trade-off between accuracy vs interpretability, bias and data deficiency, and actionability of the explanation. On the other hand, human-centered explainability challenges are associated with context sensitivity, user experience variability, and legal and ethical bias.

This dissertation advocates the user-understandable explainability of AI-enabled systems by demonstrating extensive experiments introducing explainable systems on three different application scenarios. This thesis explores how explanation types, forms, and representations can vary across different application scenarios, what important facts and rationale should be considered in generating explanations, and how the explanations can be made actionable so that users can take action for possible changes along with understanding the decisions from AI models. We propose and introduce explainable techniques for different application areas, including smart home, e-commerce, and NLP. We demonstrated explanations generated by multiple explainability techniques. Our evaluation of the efficiency of explanations in smart home applications has practical implications. We have introduced a new metric for explainable smart home applications, and the results of our study revealed that the proposed explainability techniques can efficiently identify the reasons behind predictions. The empirical user study to evaluate the generated explanations for NLP applications also highlighted what facts, rationale, and reasons should be selected in making explanations

actionable, providing valuable insights for developers and professionals in the field.

1.4 Research Challenges

Achieving human-understandable explainability for complex ML and DNN models is much more challenging due to multiple factors that include the complex black-box nature of the model, the interim trade-off between accuracy vs interpretability, user expertise variability in providing contextual explanations to help users understand decisions, representing the facts behind predictions in a human-centered way, and making explanations actionable so that users can take necessary action based on predictions and explanations. There has been a massive interest in opening the black box to achieve explainability without sacrificing accuracy. However, most of the works in this area of explainability can provide explanations that can be useful to increase models' performance. However, the explainability tools and techniques are not readily applicable to provide human-centered actionable explanations to the general users. Instead of discussing all associated challenges for human-centered explanations for AI-enabled systems, we briefly discuss the challenges considered to be addressed in this dissertation.

- **User experience variability:** Different users have different information needs to understand the decision from AI-enabled systems. On top of that, the context of applications also indicates the need for different explanation types. For example, the explanations for the smart home users for their household energy demand forecasting systems will be completely different from the explanations for the users of e-commerce websites for fake review identification.

  Existing explainability techniques might not fit the problem domain to tackle such a challenge [104]. Instead, we need to canonicalize

and contextualize the explanations to represent them so that users can understand the rationale behind the prediction. On the other hand, literacy in AI and technical applications for general users is also a considerable factor that can challenge the success of human-centered explanations. For a specific context of a single application, users might have different expertise and understandability in sense-making. Therefore, achieving human-centered explainability for AI-enabled systems is a formidable challenge.

- **Context Sensitivity:** The explanations' effectiveness to the end-users depends heavily on the context of the application [12]. Let us consider a well-known loan application example in XAI, where the loan applicant might ask for explanations from the bank official about why his/her loan has been rejected [209]. Here, the context of the explanations is from the users' side. Nevertheless, let us consider the context of the manager or the bank stockholders. They might also ask for explanations from the AI-enabled systems about the factors they should consider to get more revenue. Other than that, the expertise of different users varies widely. That is why it could be more challenging to meet the information needs in the form of explanations.

- **Bias and Data Deficiency:** Effective training of an ML model needs an adequate amount of data. Though the current computing world is enriched with high-quality available data, some applications still need enough data with proper annotation. For example, in the case of personal thermal comfort preference modeling, the datasets are not enough to train a DL model. Because it is very time-consuming and, at the same time, expensive to hire so many humans to annotate the data. Along with data deficiency, sometimes the datasets are biased and imbalanced. For example, the dataset for credit card fraud detection is extremely imbalanced because very few amounts

of transactions are fraudulent, and more than 99% of transactions are valid [287]. With such an imbalanced dataset, the learning process of any ML model could often be biased and provide outcomes biasedly. Providing explanations for such a biased model will create other trust issues. Therefore, adequate balanced data is mandatory for an ideal trained ML model.

- **One XAI technique does not fit all:** Although the current XAI-research community has made significant advancements in explainability techniques over the last decade, it still faces dilemmas such as the notion that "*one solution does not fit all* [104]." In other words, explainability heavily depends on the context of the problem at hand and the variability of technical and general user expertise. Consequently, no single explainability technique can provide explanations, even on a small scale, for all types of problems. Thus, achieving human-centered explainability is much more challenging than technical explainability alone. One approach could involve leveraging insights from technical explainability methods and presenting them in an easily understandable way for general users. Alternatively, ML models may need to involve humans in decision-making and provide explanations based on their feedback, which is even more challenging.

- **Actionable Explanations:** The generated explanations should be actionable. From the perspective of AI practitioners, insights gleaned from technical explanations can inform modifications to the model's structure or other aspects to enhance model performance. This makes the explanations actionable. Similarly, from the standpoint of end-users, human-centered explanations must fulfill a dual requirement: users should not only comprehend the decisions but also be empowered to take appropriate action based on the insights to optimize outcomes.

1.5 Research Questions

The primary goal of this dissertation is to advocate for generating human-understandable explanations for the decisions or predictions from ML models. As mentioned in the previous section, the explanation types and representations might differ based on user experience variability, context, and application discourse. Following this, we aimed to make systems explainable by providing comprehensible explanations for different applications covering different contexts and users with variable experiences. We planned to study how the explanations can be different based on the application scenarios, what the facts and priorities XAI models need to utilize to generate explanations, and how the explanations can be actionable; hence, users can use them to corrective action to make their interaction better with the systems and optimize their practice. We explore introducing XAI techniques in multiple AI-enabled innovative applications, including smart homes, NLP, and e-commerce. Overall, we conducted experiments in different application domains towards achieving human-understandable explanations by exploring the following research questions listed below:

RQ 1: ***What techniques and strategies can be employed to achieve high-performance ML models addressing technical challenges such as data imbalance, data inadequacy, and model bias?***

Since this dissertation focuses on achieving easy-to-understand explainability for different application contexts and scenarios, we first need high-performance ML models for particular tasks before explaining the prediction. In answering this research question, we strive to achieve high-performance ML models for a particular problem by introducing new effective techniques. To do so, we must address the challenges, including extreme data imbalances,

inadequacy and model bias, by following effective strategies and techniques.

RQ 2: ***How can human-understandable explainability be achieved for ML models within a specific application domain?***

Demonstrating the outcome from state-of-the-art XAI techniques in multiple real-world problems in different domains, we try to explore how we can achieve explainability and what the gaps and challenges remain to implement in different application scenarios, including smart home, e-commerce, and NLP. We explored three different applications by conducting experiments and proposing explainability methods to provide users with easy-to-understand explanations for certain decisions from the AI model. For smart home application scenarios, we propose *ForecastExplainer* to explain the prediction from a household energy demand forecasting system powered by a complex deep neural networks (DNNs) model. We observed that the explanations for the time series forecasting models are two-dimensional. We can not provide explanations just in the form of the contribution of different features. Instead, we had to consider the temporality, the time associated with the energy consumption. Unlike classification and regression tasks, the time series forecasting problem differs, and time stumps play a vital role here.

We then explore how the explanations can be generated and represented for natural language processing applications. However, we consider different applications for NLP applications, including patent classification and fake review detection tasks. We proposed high-performance prediction models and introduced explainability methods for explanations for the users to make them understand why a particular patent is classified to a specific class or why a review is predicted as fake, which are provided in the

forms of highlighted words. We incorporate a layer-wise relevance propagation technique to identify the weight of every word that refers to the importance of a particular class. Then, different visualizations are applied to represent the explanations in the most accessible forms. Along with the user-centric applications, we also introduce the explainability method for product back-order prediction tasks, a prominent problem in e-commerce, for the company's stakeholders so that they can take necessary steps to get rid of being a product back-ordered. For the fake review identification task, the empirical evaluation of generated explanations with human subjects demonstrated what information needs to be considered to generate and represent explanations.

RQ 3: ***How do explanations vary across different real-world applications?***

The types and formats of explanations significantly depend on the context of the applications and even on the application scenarios. For example, technically, explanations for energy demand forecasting systems differ significantly from those for patent classification systems. In the former, explanations represent the energy consumption for different appliances and corresponding times. In contrast, in the latter, explanations mainly depend on scientific terms and innovations related to specific research fields (such as electricity patents or chemistry patents).

Moreover, the granularity of explanations also varies within different application contexts. While end-users of one application may require detailed, fine-grained, and easily understandable explanations, others may expect higher-level explanations specific to the context. In certain applications such as medical diagnostic systems, it is imperative to consider ethical and regulatory requirements when presenting explanations to users. Building on

these observations, we aim to explore how the types and formats of explanations vary across different applications. Therefore, we conducted experiments introducing various forms of explainability using state-of-the-art XAI techniques in smart homes, e-commerce, and NLP application scenarios.

RQ 4: ***What underlying facts and rationale should we consider when generating explanations within application scenarios?***

To achieve human-understandable explainability, it is crucial to consider the underlying facts, context, and rationale before representing explanations based on the model's priorities. To understand context, we should employ facts and rationale when generating explanations. We demonstrated explainability in multiple application scenarios and investigated the significant requirements that must be considered. The most critical underlying facts and rationale include end-user expertise, relevance to the specific task at hand, and the actionability of the explanations. To explore this research question better, we conducted a wide range of experiments in different types of applications, which are also diverse in terms of ML tasks. The ML task included multivariate time series forecasting, classification, text processing, and multi-level classification. Generating explanations for such diverse applications, we tried to explore what underlying facts and rationale are essential to consider and represent in explanations.

In the context of NLP applications, for example, the users of explainable patent classification systems expect the provided explanations not only to make sense of why a particular ML system classifies a patent to a specific class but also to help them by providing holistic ideas on how they should present text information in future patents to be classified to their desired patent class.

RQ 5: ***How can actionable explanations be generated to optimize practice for given application contexts?***

Usually, explanations help users understand decisions made by AI-enabled systems. We investigated how explanations can be generated by representing facts and rationale to help users understand the decision so that they might take the necessary action to optimize their interaction with intelligent systems. However, when explanations are actionable, stakeholders can take appropriate actions to optimize outcomes in various application scenarios. For instance, in a product backorder prediction scenario, stakeholders expect to be informed of an impending backorder and get the reasons provided by actionable explanations so that they can make specific changes to prevent the backorder, maximizing revenue and minimizing loss.

Similarly, in the context of smart home applications, providing consumption scenarios for future energy usage in terms of activities rather than appliance-level consumption enables users to optimize their consumption habits across different activities. Our research findings support the generation of actionable explanations that represent activity-based consumption, benefiting end-users in understanding decisions and taking appropriate actions to maximize outcomes. Additionally, the granularity of explanations varies across applications, with some users requiring detailed, fine-grained explanations while others need context-specific high-level explanations. Furthermore, ethical and regulatory considerations must be addressed when presenting explanations to users in applications such as medical diagnostics systems.

1.6 Structure of the Dissertation

This dissertation consists of five different parts.

Part **I** presents the introduction & overview, related work and research design & methodology. This part consists of three chapters. Chapter 1 presents the motivation, associated research challenges, research questions and the structure of the thesis. We discuss the related works in XAI, human centered XAI with the possible research gaps in chapter 2. Chapter 3 presents the details about the application area selection, technical highlight of applied methods, and study outline.

The following three parts discuss our proposed methods and findings on introducing XAI in three research areas. Part **II** focuses on explainable smart home applications. Chapter 5 argues the need for human-centered explainable systems in smart home environments. Chapter 6 presents an explainable energy demand forecasting system. The personal thermal comfort prediction system is introduced in chapter 7. Part **III** is about explainable business applications. Chapter 9 presents our proposed explainable product backorder prediction system and discusses the findings from generated explanations.

Next, in part **IV**, explainable methods on two different NLP applications have been introduced. In chapter 11, we present an interpretable patent classification system to generate explanations to understand the predicted class. Chapter 12 presents an interpretable fake review detection method and the user evaluation of the generated explanations.

In the last part (Part **V**), we discuss the findings from the experiments and analysis of the above three parts and the conclusive remark with future work. Chapter 13 discusses the findings from the introduced explainable techniques in three different application scenarios. At last, chapter 14 concludes the dissertation with some future work to address the potential limitations.

2 Related Work

In the last decade, there has been an enormous interest in the AI research community in explaining ML models and their predictions to have more interpretable, transparent, and trustworthy AI applications. This chapter presents an overview of state-of-the-art XAI methods, the progress in human-centered XAI, and the related research works on ML and XAI techniques in the application areas studied in this dissertation, including smart home, business, and NLP.

2.1 Explainable Artificial Intelligence

To define XAI, we quote the definition from the International Business Machines Corporation (IBM), one of the pioneering institutions in explaining ML models, as follows:

Definition 1:: "*Explainable artificial intelligence (XAI) is a set of processes and methods that allows human users to comprehend and trust the results and output created by machine learning algorithms.*[2]"

State-of-the-art XAI techniques can be broadly classified into different criteria, based on scope, model types, complexity of the techniques, and methodology used [12]. A key factor in these techniques is feature importance. Techniques that are based on feature importance can be referred to as scope-based XAI techniques. In such techniques, the features of data samples are used to explain the overall model priorities and individual predictions, representing which features have positive (and negative) contributions towards the model's overall decision-making or specific decisions.

XAI techniques that focus on explaining overall model priorities are known as *global*. In contrast, XAI methods that focus on explaining

[2]Source: `https://www.ibm.com/topics/explainable-ai`

individual decisions from the model are called *local* [4, 277]. One of the most prominent global XAI techniques is Shapley Additive Explanations (SHAP), proposed by Lundberg et al. [195] based on game theory. Their hypothesis is similar to identifying individual players' contribution towards a game's outcome. SHAP tries to explain decisions in terms of feature importance (both negative and positive directions toward the decision), where features are treated as players in a particular game (i.e., the application area), and the prediction is treated as the game's outcome.

Local Interpretable Model-agnostic Explanations (LIME) [244] can be considered the most widely used local XAI technique, which explains specific predictions from the models. LIME trains a surrogate interpretable model that mimics or approximates the performance of the complex model. Then, the surrogate model is employed to explain particular decisions from the original model [244].

Generally, there is a trade-off between the complexity and interpretability of ML models. Less complex methods are supposed to be more interpretable. The accuracy of ML models is inversely proportionate to the degree of interpretability. Ideally, the ML model should be highly accurate and interpretable simultaneously. However, this is practically impossible due to the complexity of the models. Usually, complex models such as deep neural networks have thousands or millions of parameters related to predicting a particular decision. These complex models have high accuracy but are also less interpretable because of their complexity.

On the other hand, basic logistic regression models offer high interpretability. Their simplicity makes it easy to uncover any model's decision-making process. However, this high level of interpretability comes at the cost of accuracy. In Fig. 1 and 2, we present the relationship based on the accuracy and the interpretability. We can see from both figures that, the DNNs model is highly accurate but has very less

interpretability. The next accurate ML models are ensemble methods such as XGBoost and random forest, followed by kernel-based methods, clustering, k-nearest neighbors, and decision trees. However, the relationship based on the interpretability is inverse.

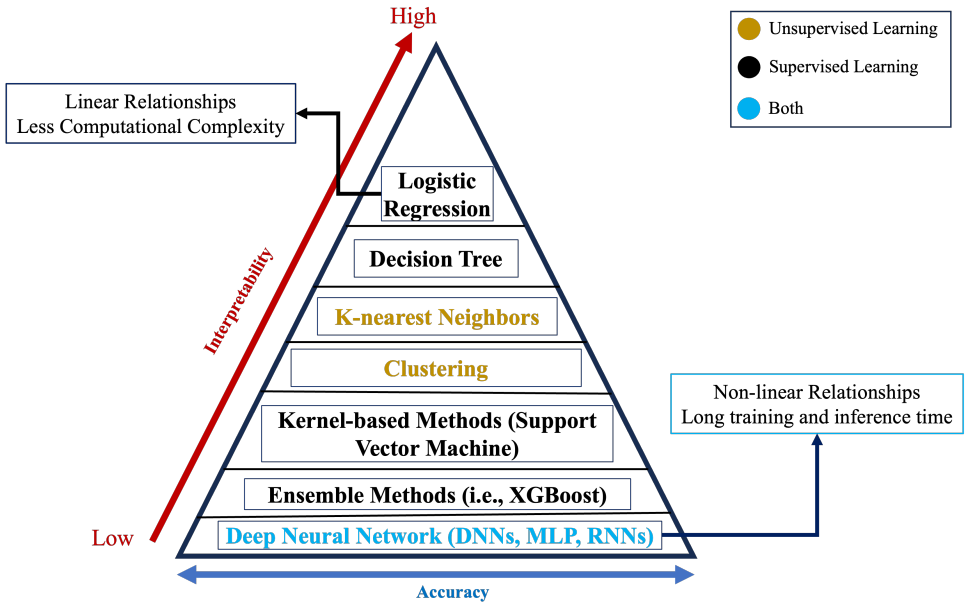

Figure 1: Accuracy vs Interpretability trade-off in machine learning models.

In a practical scenario, we try to have a model that does not sacrifice considerable accuracy but will achieve high interpretability. However, the model selection for a particular application depends on many factors, including the users and the objective of the application. To interpret the prediction from the complexity of highly accurate models (i.e., DNNs and Ensemble models) is possible thanks to the recent advancement in uncovering the decision-making process.

Based on the complexity of the models, we can categorize XAI techniques as intrinsic and post-hoc. The design and implementation of intrinsic models ensure that the predicted decisions are inherently interpretable. There is no need for other explainability techniques to comprehend decisions from such models. However, more complex ML mod-

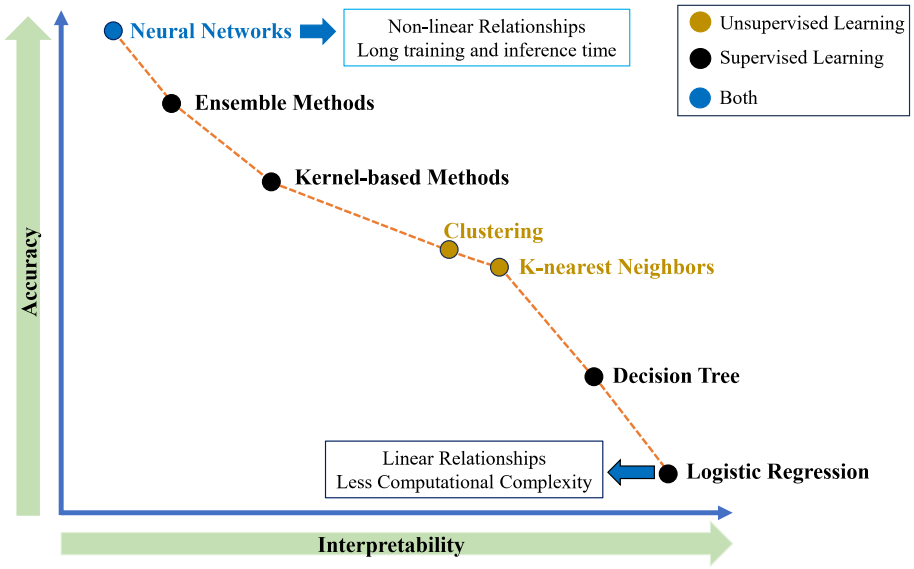

Figure 2: Accuracy vs Interpretability trade-off in machine learning models in terms of line chart.

els like DNNs and autoencoders need another explainability model on top of the ML models, referred to as post-hoc explainability techniques. Model-agnostic XAI techniques offer post-hoc explainability and can be applied to complex models to explain their decisions.

Some explainability techniques can be applied to any ML model, while others are designed only for specific types of models. XAI techniques that apply to multiple types of models to explain decisions are called model-agnostic XAI techniques. SHAP and LIME are both model-agnostic techniques and can be applied to classical ML models, from decision trees to support vector machines, and DL models, from deep neural networks (DNNs) to recurrent neural networks (RNNs) and autoencoders [277, 4, 150].

Based on the methodologies used, there are two major categories of XAI techniques: back-propagation-based and perturbation-based methods.

XAI techniques based on back-propagation propagate the model's output backward from the output layers to the input layers of the DNNs. In the back-propagation phase, XAI techniques try to capture the important changes in different nodes in every layer and map the weight changes to compute the contributions of specific nodes. After the backward propagation phase ends, these methods can identify the contribution of different input features towards the predicted output. These can also be called gradient-based techniques. Major examples of these kinds of methods are saliency maps [295], layer-wise relevance propagation [19, 211], and Deep Learning Important FeaTures (DeepLIFT) [293].

To explain the prediction from a DL model, Montavon et al. [211] proposed layer-wise relevance propagation (LRP)-based XAI technique. LRP redistributes the weights of a particular neuron from the output layer to the input layer through the intermediate layers of DNNs. By this process, it assigns a weight to each node (neuron) in every layer by weight redistribution and eventually marks different input features and assigns weights that have the influence or contributions (both positive and negative) towards the predictions [211, 285]. Deep Learning Important FeaTures (DeepLIFT) has been introduced by approximating the Shapley values to decompose the deep neural network and explain particular predictions from DNN-based models [293]. Grad-CAM, another notable XAI technique, has been proposed based on gradient localization, which can explain predictions for image processing applications.

Example-based explanations can make sense of why the ML model predicted a specific outcome. These explanations provide similar examples with similar feature values contributing to the same prediction [156, 117]. Contrary to example-based explanations, contrastive, counterfactual, and *what-if* explanations also play a vital role in understanding [310, 213, 315]. Counterfactual explanations can make people understand by changing specific feature values that might overturn the

prediction so that users can comprehend the particular decision [315]. *What-if* scenario-based explanations demonstrate what will happen if the user changes one or more specific values of the samples [309].

Several XAI techniques are introduced for AI experts based on subspace explanations and probabilistic interpretation for Bayesian decision trees [172, 263]. These explanations are also helpful in understanding the models' predictions. Other notable XAI techniques include individual conditional expectation (ICE) [45], diverse counterfactual explanations (DiCE) [213], and partial dependency plots (PDP) [110], which have also been introduced following different hypotheses to explain the predictions of ML models. Combining different primary XAI techniques, some notable XAI frameworks can be used as a library to apply for explaining individual ML model decisions, including Captum [166], AIX360 [24], and Anchor [246].

2.2 Progress Towards Human-centered XAI

The significant progress in explaining the decisions from complex ML models is centered on AI practitioners and developers to debug and improve the performance of the models by modifying parameters and architectures after careful analysis of the provided technical explanations [250]. Nevertheless, many studies have been conducted to evaluate the current XAI techniques from a human-centered perspective. Different studies were also carried out on eliciting the requirements for achieving human understandable explainability with user studies.

One of the primary goals of XAI is to make people understand the predictions from AI models. However, there is a clear difference between the user's perspectives on comprehending the explanations and how they are designed and presented using underlying facts and rationale [67, 122]. Rong et al. [251] classified the understanding of explana-

tions into objective and subjective. Objective understanding is related to comprehending the decisions of the models in the application domain, and for evaluating this, a proxy application with a questionnaire can be a good tool. On the other hand, subject understanding can depend on the user's perspective, and it generally can be evaluated through user study after the task has been completed [251]. A concept-based explanations framework has been introduced by Ghorbani et al. [107] where the authors proposed principles to overcome the problem with feature importance-based explanations. Several studies have been conducted to evaluate the users' understanding of the explanations in different application areas, including finance [48, 2], education [38] and image processing [47].

The investigation on how the non-technical users are capable of understanding the decision-making priorities of the model by providing global explanations has been conducted by Chromik et al. [67], and they concluded that the global explanations in terms of features importance are not enough. The effectiveness of global model-agnostic explanations generated by SHAP for applications in finance and education has been evaluated by the general users [38]. Their findings also indicated that global explanations are insufficient in understanding model behavior [280, 38].

Besides users' understandability of the explanations, trust in XAI systems plays an essential role in the real-world adoption of AI models. Moreover, it is a popular argument that XAI systems can increase user trust [150, 38]; hence, human-centered explanations must be trustworthy. Studies across different research domains, including medical [229, 308], self-driving autonomous vehicles [70], recommendation systems [187, 227] and banking applications [266] has been conducted to measure whether the explanations increase trust on the automatic decisions.

From the human-computer interaction (HCI) perspective, the usability of XAI systems should consider different concepts such as ease of use, helpfulness, and ability to detect undesired decisions from the systems. To measure the users' satisfaction level on XAI-enabled systems and their generated explanations, researchers introduced "Satisfaction Scale" considering multiple aspects [229, 126, 48]. Based on the users' ratings on the generated explanations, Nourani et al. [223] measures how helpful the explanations are in comprehending the decisions. Similarly, Zhang et al. [340] and Buccinca et al. [50] investigated to measure the helpfulness of the explainable systems. To evaluate whether the generated explanations can audit the models, multiple studies conducted user evaluation considering the fairness of the decisions [265, 238], model bias [298, 243] in decision-making and undesired decisions [159]. Experiments based on simulateability also concluded that current interpretability methods cannot explain model behavior [189, 251].

The explanations generated from prominent methods, including LIME and anchor, have also been evaluated by Hase et al. [122]. They found that counterfactual explanations perform better than others in understanding, and LIME explanations were practical in a few cases. Abdul et al. [2] conducted an empirical user study with more than 80 participants to measure the effectiveness and user satisfaction of the generated explanations. In summary, in case of understanding perspective, the empirical user studies on multiple real-world applications suggest the contradiction in achieving the XAI goals [188, 30, 235, 317].

2.3 XAI in Applications

This section provides a brief overview of the relevant AI and XAI research works on different applications related to the research areas investigated in this dissertation. We focus on three different research areas: smart home, business, and NLP applications.

2.3.1 Smart Homes Applications

The first application area of this dissertation is the smart home, which includes applications for explainable energy demand forecasting and predicting thermal comfort preferences for automatic heating and cooling systems. Here, we discuss notable research related to smart home applications.

The applications of AI-enabled systems are now becoming popular in smart home scenarios [161, 162, 163, 311, 278]. AI-based predictive models and recommender systems are utilized in the home environment to provide more comfortable living space to the inhabitants [1]. Moreover, it intends to provide feedback to the inhabitants to be more optimized in consuming energy and hence try to decrease carbon footprint and households' energy-related costs [77].

Energy demand forecasting is a kind of smart home application that can provide the inhabitants about how much energy they will need in the upcoming week or month [164, 163, 278, 153]. With overall future energy demand forecasting, it can also provide appliance level prediction so that the household member can understand how they might optimize their energy consumption. Another impressive example of an AI-enable system in a smart home is an intelligent heating and cooling system based on the prediction of occupants' thermal comfort preferences [25, 139, 177, 180]. In other words, the system can change the house's or workplace's temperature after predicting the thermal comfort preferences. Hence, the outcome after the AI-actuator is that changing the temperature provide occupants with a more comfortable and pleasant environment.

Forecasting household energy demand to make people aware of their energy consumption is a multi-variate time series forecasting task. Making multi-variate time series forecasting explainable is challenging com-

pared to classical classification and regression tasks. Because explanations of mufti-variate time series forecasting models related not only to the features of the sample but associated with the corresponding time frame [278]. Therefore, the explanations are generally two-dimensional. The existing XAI techniques often can not be used directly to explain the time series forecasting task.

Several survey papers on explainable techniques for time series data have discussed the overview and existing methods [249, 89, 32, 264, 260, 255]. An interpretable boosted regression model is introduced by Ilic et al. [133] where they generate explanations for predictions employing regression tree. In terms of the saliency map, an explainable technique proposed that can decompose and explain the prediction from convolutional neural networks (CNNs) model [256, 257]. To explain the prediction from the auto-encoder-based energy demand forecasting model, Kim et al. [163, 162] proposed an interpretable method that can explain the prediction in terms of the heatmap.

Grimaldo et al. [111] introduced visually interactive and explainable energy demand analysis systems incorporating different prosumer scenarios. Jiao et al. [142] experimented on classical ML models to identify the critical factors for energy demand forecasting in residential buildings. Shapley value-based explanation technique is developed to interpret the decision from DNNs-based energy forecasting model [232]. The two most widely used XAI techniques, SHAP and LIME, have also been used in energy demand classification task [40].

Research on predicting personal thermal comfort prediction to provide a more pleasant and comfortable indoor environment by changing temperature and other parameters has got a significant interest in recent time [334, 78, 329, 328]. Generally, ML models for thermal comfort preference prediction depend on high-dimensional feature spaces, including physiological, environmental, and weather data. There are two technical

challenges associated with thermal comfort preference modeling, including inadequate data samples to train the models and high-dimensional features space that might mislead the models [7, 237, 334, 78]. To address the data inadequacy challenges, researchers employ different synthetic data generation techniques such as generative adversarial networks (GANs) and their variants [237, 334, 78, 237]. To identify relevant features and filter out irrelevant and correlated features, feature selection techniques have been applied as a first step before training the model [152, 289, 194].

2.3.2 Business Intelligence Applications

Predictive decision analysis and modeling are largely employed in e-commerce, banking, and other business domains. Some examples of AI-enable systems can be customer churn prediction, product backorder prediction and fraud detection, inventory management, and supply chain optimization [128, 286, 119, 204]. Identification of fake or AI-generated reviews in e-commerce platforms is now evident by applying advanced NLP techniques so that fake or artificial reviews of products can not mislead customers. In summary, the customer and company stakeholders can benefit from such systems, where general users might get help to become more aware of fraud. The company can make more profit and avoid possible losses by taking corrective action based on predictions from the AI model.

The applications in business intelligence can be broadly categorized into two classes based on employed AI models. The first category is the type of applications that employ classical ML models, including decision tree [119], support vector machine (SVM) [119, 226], and K-nearest neighbor. Another category employed sophisticated DL techniques, including DNNs, RNNs, and GAN [185, 286, 174]. The use of XAI techniques in business applications is also evident in recent times [226, 225].

However, the remainder of this subsection presents a brief overview of applications of ML models and the use of XAI techniques in the business area, especially in product backorder prediction.

One of the prior methods of product backorder prediction is proposed by De et al. [79], where they overcome the imbalance problem with the combination of oversampling and undersampling techniques. Ntakolia et al. [226, 225] proposed two different methods to have an explainable inventory management system by applying a global XAI technique, Shapley values. They relied on classical ML models for modeling product backorder data, including SVM, random forest, and XGBoost. Hajek et al. [119] introduced profit-maximization techniques by training ML models for predicting product backorder. They have yet to explore the explainability techniques. Islam et al. [134] used trained random forest and gradient boosting model to predict backorder for supply chain management system.

Several researches on product backorder prediction introduced DL models [185, 262, 174, 286]. Lawal et al. [174] applied an RNN architecture to model product backorder after applying preprocessing with a min-max scaler and addressing the imbalance problem with the ADASYN oversampling technique. Deep neural networks have also been proposed for identifying product backorder [286]. Saraogi et al.[262] applied an un-supervised approach with auto-encoder to predict future backorder. Li [185] applied a series of ML models in his PhD thesis for modeling product backorders.

However, almost all methods mentioned above employed complex ML models for backorder prediction, but except few [226, 225] studies, none of them try to explain the predictions from the trained model. Though Ntkolia et al. only explore global explanations for the overall models' priorities. One of the reasons why we chose this business task is to make the predictions explainable so that the stakeholders, i.e., the inventory

manager, can take corrective action to decrease future loss.

2.3.3 NLP Applications

There has been significant interest in explaining the predictions from NLP models, including explainable sentiment classification [345, 6, 336], and interpretable hate speech detection [149, 199]. Explainable techniques including SHAP [195], LIME [244], ELI5[3], anchor [246] and LRP [211] are the most notable methods employed in explaining decisions from text classification models. In this thesis, we chose two NLP applications to investigate, including patent classification and fake review identification tasks. Following a brief discussion on the breakthrough in NLP tasks, we briefly overview the related works in the two tasks mentioned above.

The revolutionary methods of representing text by introducing word-embedding models have changed the dimension of NLP applications and achieved much higher accuracy in almost all sectors. Compared to the classical text representation techniques such as bag-of-words (BoW) and term Frequency - Inverse Document Frequency (TF-IDF), the semantic representation of words can capture much better semantic information for a longer context. Text-embedding models including word2vec [175], sentence2vec [203], Glove [233] and fasttext [46] have been introduced to represent text in high dimensional semantic vectors for modeling NLP tasks. Therefore, the NLP task, including sentiment analysis, textual similarity estimation, question-answering, and machine translation, achieved higher performance than ever. However, after introducing transformer-based text representation techniques, the NLP applications achieve new state-of-the-art performance, which, in some cases, is higher than that of humans. Transformer models from BERT [81] with its varient such as DistilBERT [261], RoBERTa [192] to XLNET [333],

[3]ELI5: `https://github.com/eli5-org/eli5`

Electra [68] are GPT [101] performing better to model NLP tasks. Recently, large language models have been introduced, which have surpassed the performance of all previous methods in language generation and understanding tasks. This dissertation investigated two NLP applications, including patent classification and fake review identification, and introduced XAI-enabled models with easy-to-understand explanations.

The number of patent applications in the last decade has increased exponentially. It needs a considerable effort to classify them manually for the patent experts [288, 183, 178]. The preliminary methods for classifying patents relied on applying BoW and DF-IDF representation with classical ML models, which included decision trees and SVM. Those text representations have severe limitations in that they can not capture semantic and contextual information of the texts. Multiple methods have been proposed to classify patents employing DL techniques, including CNNs, LSTM, BiLSTM, and hybrid approaches such as CNN-LSTM and CNN-BiLSTM [183, 178, 63]. To achieve a multi-level patent classification system, researchers applied pre-trained word-embedding [236] and transformer models including Glove, fasttext, BERT, XLNet, RoBERTa and Electra [68, 252, 141, 148].

Similar models are also applied in fake review identification tasks using text representation methods, including pre-trained word-embedding and transformer models. The classification models include basic ML models to complex and hybrid DL models [85, 335, 231, 230, 296, 80]. [80]. Duma et al. [85] proposed a deep fake review detection method by analyzing ratings and latent text feature. Introducing features such as authenticity and analytical thinking, Alsubari et al. [14] applied an RNNs-based deceptive review identification method. The sentiment of the reviews is also considered to classify whether the review is fake or original [294]. Topic modeling and semi-supervised GAN have been

employed with an attention mechanism to model fake review detection tasks [43, 144]. Mohawesh et al. [207] proposed several deep learning models that include Bi-LSTM and CNN, and they also attempted to explain the prediction with SHAP. Previously, they also considered addressing concept drift for identifying fake reviews [208, 206].

However, neither selected NLP task in this thesis has been studied well enough to explain the predictions for the general users. For the patent classification task, it is significant to explain to the patent experts why the ML models predicted a patent text to one of the hundreds of classes. The same holds for users of e-commerce platforms, where users might be fooled by fake reviews posted by retailers or their competitors. It needs to be detected automatically, providing explanations so that users can trust the systems.

3 Research Design & Methodology

This dissertation is focused on introducing explainable models and demonstrating their applications in various areas. It investigates how explanations for understanding decisions from complex AI models vary across different application scenarios, contexts, and user experience variability. The research delves into three diverse application areas: smart home, NLP, and business intelligence.

3.1 Selection of Application Areas

It is of utmost importance to select application areas for in-depth exploration and investigation to address the challenges and answer the research questions. We have chosen research areas based on several criteria directly linked to the research questions. Our focus is not just on understanding how explanations vary across different application domains but also on the potential impact of this understanding. We aim to identify the facts and rationale that should be considered when generating explanations for a given application's context and to produce actionable explanations that can guide corrective actions, thereby making a significant contribution to the application at hand.

Our objectives are underpinned by the recognition of significant research challenges. These include the delicate balance between accuracy and interpretability, the variability of user experience, the sensitivity to context and bias, and the issue of data deficiency. These challenges are not obstacles but rather opportunities for us to delve deeper into our research questions. They guide us in selecting the three application areas in multiple sub-tasks are diverse and can be investigated to find the answers to our research questions and overcome the challenges to achieve explainability.

We also stress the importance of investigating explainability techniques

across various ML tasks, including time-series forecasting, tabular classification, and text classification. In the following, we discuss why the three application areas were selected and how they provide us with the platform to investigate the research questions, serving as the key areas where our research will be applied and tested.

1. **Smart Home:** First, we selected the smart home application area to investigate the explainability methods to generate explanations so that the general users might understand the complex decisions from AI models. Smart home users are laypeople who might need a basic understanding of how systems make decisions. We found that smart home applications are understudied in the direction of XAI. Some ML-related works have been related to energy demand forecasting and thermal comfort modeling. However, they all are predictive systems applying black-box ML models. Few studies in energy demand forecasting tried to explain the models' decisions; however, those methods aimed at using technical explainability to debug and improve the model's performance.

   Energy demand forecasting is a multivariate time series prediction task that is not similar to traditional classification task. The predictions depend on the consumption of different appliances and the associated time. Hence, the explanations are also two-dimensional. Therefore, generating an explanation mapping the contributions of features corresponding to the time is challenging. On top of that, it should be easy to understand since the users are lay people.

   Significant attention has been paid to modeling the thermal comfort preference inside household and corporate buildings, applying ML models to provide a comfortable environment, and controlling the indoor temperature based on the preference prediction. This area is challenging for two different reasons. First, it has high-

dimensional features, which makes it challenging to select the most important and which ones should be filtered out. Generating explanations for high-dimensional datasets is much more challenging than for smaller dimensions.

2. **Business:** We selected one business application, product backorder prediction to have diversity in application areas. The explainability of such applications has yet to be studied in the literature. Usually, product backorder prediction is a very challenging task for a few reasons. One of the most important reasons is that product backorder is rare but essential even in inventory management. This rarity leads to a highly imbalanced dataset, which hinders the performance of ML models and leads to bias. In addition, the stakeholders of such applications are not general users but the company owner and inventory manager. Given this context, explaining the decisions from the ML models to make stakeholders understand the reasons is a formidable task due to the data imbalance and user experience variability. The stakeholders here would not only like to understand the decision but also want to overturn it by taking corrective actions.

3. **Natural Language Processing:** In the above application areas, we investigate to explain the decisions from ML models trained on datasets with time series and tabular numerical data. Since two of our research questions were related to studying how the explanations varied across application areas and what facts and rationale needed to be considered in explaining the decisions, we would like to investigate the decisions from NLP applications. Unlike the other two selected application areas, the data in NLP areas is a text. So, in terms of dataset types, the application domains will be diverse.

   Hence, we chose two NLP applications: patent classification and fake review identifications. These applications were chosen be-

cause they represent different user groups and decision-making processes, which allows us to explore the generalizability of our findings. The two applications are different in terms of the users. However, both applications are understudied in providing explanations for the general users. The users of patent classification systems are patent experts who manually classify patents into different classes, considering the scientific contributions and scope of the patent text. On the other hand, for a fake review identification system, the users are general people who want to know whether the review is fake or original.

3.2 Technical Highlights

The overview of selected application areas with the particular sub-tasks is depicted in Fig. 3, which also highlights that explainability is the central focus of this dissertation. We investigated explainability methods on diverse applications of different ML tasks, including tabular classification, multi-variate time-series forecasting, and text classification. The datasets are collected from different sources that contain time-series household energy consumption datasets, thermal comfort datasets with physiological, weather, and environmental features, inventory datasets, and text datasets for patent and fake review identification tasks. The types, forms, and representations of explanations widely varied across those applications.

Therefore, we introduced different ML or DL methods selected after careful analysis and preliminary investigations for every sub-task. Considering the dataset, users experience variability, and application context, the explainability methods are selected, and the explanations are represented. However, we broadly relied on model-agnostic explainability methods since these can explain the decisions of different models. Here, we present the highlights of the methodologies applied to different ap-

plication domains.

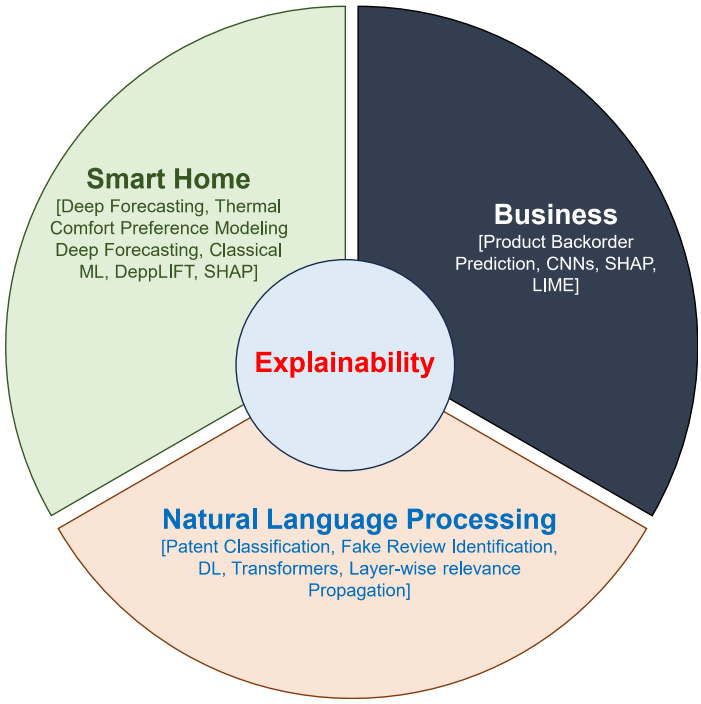

Figure 3: The application areas investigated in this dissertation with the objective of achieving explainability

We first propose a deep energy demand forecasting model based on recurrent neural networks in the smart home application context. To decompose the decision-making procedure of the forecasting model, we use DeepLIFT explainability tools and introduce them to explain the forecasting. We also map the contributions of appliances with the corresponding time so that the explanations represent the facts, including features and time. An evaluation metric is also proposed to measure the efficiency of the explanation by calculating the correlation between explanations and ground truth.

Our research into predicting personal thermal comfort preferences involves working with a high-dimensional dataset. To ensure the accuracy of our predictions, we conduct a comprehensive analysis of the features,

investigating whether there are any correlated and redundant features. Based on the outcome of our correlation analysis, we introduce a range of supervised feature selection techniques to filter out redundant features. To address the challenges of data inadequacy, we introduce the conditional tabular GAN to generate synthetic samples. Finally, we train several classical ML models to identify preferences and explain the global priorities of the model using Shapley values.

We propose convolutional neural networks-based predictive models for the product backorder prediction task. Before that, we address the data imbalance problem by using ADASYN. SHAP and LIME are employed to explain the prediction of the CNNs-based backorder prediction model. SHAP has been applied to provide global explanations to understand the overall model's decision-making priorities. SHAP and LIME have been employed for local explanations for particular predictions, and the explanations are represented in different forms.

Finally, for NLP applications, we explain the predictions from DL models applying the layer-wise relevance propagation (LRP) technique. We proposed several DL models for patent classification systems, including BiLSTM and CNN-BiLSTM. On the other hand, we employed transformer models to have the semantic representations of the text used in the DL-based fake review identification system. For both applications, we redistributed the prediction from the output layer to the input layers through intermediate layers of the DL models. Finally, we present the explanations using heatmaps and word clouds. We conducted an empirical user study to evaluate the performance of the explanations for the fake review identification system with human subjects.

3.3 Study outline

We have structured this thesis into five parts. In addition to the first (Part I) and last part (Part V), we have divided the main contribution of this dissertation into three distinct parts: explainable models in smart homes (Part II), business (Part III), and NLP (Part IV). Fig. 4 provides a comprehensive overview of the different chapters in each part of this dissertation, engaging you in the process of understanding.

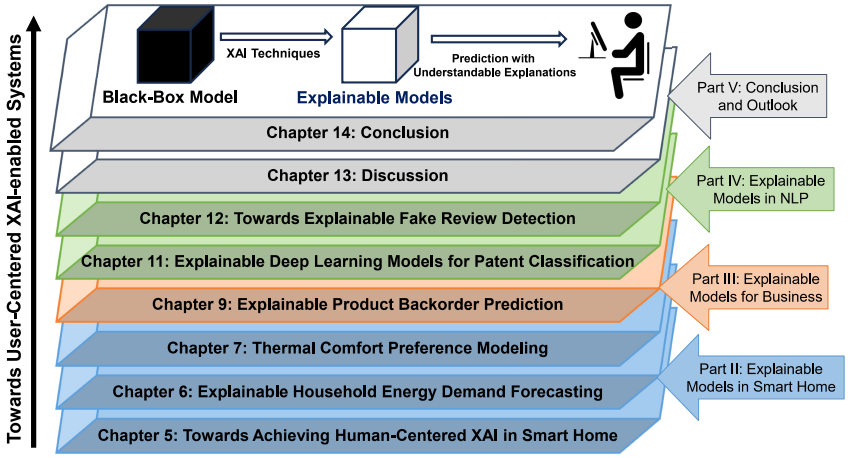

Figure 4: An overview of the dissertation in terms of bottom-up layout of different chapters

Part II Explainable models in Smart Home: We investigated the explainability in different perspectives and applications for the smart home environment. We proposed an explainable energy demand forecasting system by approximating shapely values with DeepLIFT and representing explanations mapping the feature contributions and associated time. For another smart home application, we proposed applying a generative adversarial network (GAN) to overcome the data deficiency in modeling personal thermal comfort preferences, which can be used for automatic heating and cooling control systems. Additionally, we explain the global model priorities in modeling personal thermal comfort preferences. We introduced global explanations to understand how the thermal comfort

preference prediction model makes decisions.

Chapter 6 presents an explainable energy demand forecasting system in smart homes, where a novel explainable framework is proposed to generate human-understandable explanations applying DeepLIFT that approximates shapley values. We first implemented a deep energy demand forecasting model and then introduced the explainable model to uncover the reasons behind any specific predictions from the model. We also proposed a new evaluation metric to measure the performance of our explainability techniques for the time series forecasting model. The detailed experimental results on several households' energy consumption datasets are then presented to demonstrate the effectiveness of our explainable forecasting model.

In chapter 7, we present an improved personal thermal comfort preference prediction system, that can be employed in smart home automatic heating and cooling systems. We proposed a generative adversarial network (GAN)-based technique that addressed the data deficiency challenge by generating new synthetic data samples to effectively train the thermal comfort preference prediction model. On top of that, we introduced multiple supervised feature selection techniques to filter out the irrelevant and noisy features. We present a wide range of experiment results to demonstrate the superiority of our proposed method. We then discuss the global explainability of the model by highlighting the model priorities in predicting personal thermal comfort preference.

Part III Explainable Models in Business: As noted earlier, the explanations might vary in application scenarios and contexts. We explored a different application in the inventory management system called product back-order prediction. Back-orders are orders that customers place for products that are not in stock. We proposed an explainable product back-order prediction system applying prominent SHAP and LIME explainability techniques. Unlike the smart home application scenarios,

the stakeholders and users in this application are not general people but the company owners and managers. We generated explanations highlighting the essential critical features so that the manager could take necessary action to overturn the product's back-order scenarios, eventually decreasing the company's loss.

We present how explainability techniques can be applied to e-commerce and business applications. We present an explainable convolutional neural networks (CNNs)-based product back-order prediction model, by which the inventory manager can know what product will be back-ordered (Chapter 9). We introduced SHAP and LIME, two prominent explainability methods to offer extensive explanations for the company's stakeholders so that they can take necessary steps to overturn the decision by incorporating the explanations. This chapter concluded with the experimental findings applying our methods to a real-world dataset.

Part IV Explainable Models in NLP Applications: We introduced an LRP-based explainable deep patent classification model that can explain the particular decision of the model by highlighting the related scientific words that are directly related to the particular class (i.e., electricity patent). In this case, the users are specialists responsible for assigning a particular class to submit patents. For them, comprehending the generated explanations is way easier than for lay people like smart home users. In another NLP application, we investigate an explainable fake review identification method where our model can identify whether a particular product review is real or fake (written by LLM). Here, the observation from the actual subject after demonstrating the generated explanations shows that most subjects can not find the generated explanations for the prediction, which helps them understand the AI decision. Hence, the explainability technique here failed to make the general user understand the decisions from the model.

This part focuses on the prospectus and findings of explainability tech-

niques in different NLP applications, including patent classification and fake review identification. We demonstrated how explainability techniques can be applied in different application scenarios. We demonstrated several deep learning and transformer-based text classification techniques. We also integrated LRP explainability techniques to explain the predictions.

Chapter 11 presents an explainable patent classification system. We proposed several deep learning-based patent classification techniques, and our methods' performance is significantly better than the existing approaches. We introduced layer-wise relevance propagation (LRP) techniques to redistribute the weight from the output layer of the proposed deep neural network model to the input layer through the hidden layers to assign contribution weight for each feature (i.e., words). We then assign weights for relevant words corresponding to the predicted class so that the patent examiner can understand why that particular patent is classified to a particular class. Difference visualizations of the explanations are demonstrated for each prediction.

Chapter 12 is about identifying fake reviews and presenting why the model thinks a particular review is fake. In this chapter, we proposed advanced transformer-based fake review identification methods applying DistilBERT and XLNet transformers and then modified LRP explainability techniques to demonstrate the explanations for each prediction with heatmap and word cloud-based representation. We also evaluate the explainability techniques through a user study showing the explanations generated by the LRP technique. The findings suggest that the explanations are not fully human-centered to comprehend the decision from the AI model.

Part V Research Outcome and Conclusion: This part details the discussion on the overall findings from the experiments in the chapters mentioned above. It also discusses the limitations and shortcomings of

our introduced methods. Chapter 13 presents the discussion by revisiting the research questions. The summary of the dissertation, limitations, and possible future works are presented in chapter 14.

Part II

Explainable Models in Smart Home

4 Introduction

This part of the thesis focuses on investigating the explainable models of different smart home applications. We investigated two different smart home applications, including energy demand forecasting system and thermal comfort preference prediction.

In chapter 5, we first argue on generating human-centered explanations by demonstrating technical explanations on two smart home applications with state-of-the-art explainability techniques. We then elicit several challenges that need to be addressed in achieving human-centered explainability in the context of smart homes. We conclude the chapter by highlighting the possible human-computer interaction techniques to achieve human-centered explainability in smart homes.

In Chapter 6, we then investigate how we can generate explanations for energy demand forecasting systems. For doing so, we propose a new explainable deep learning-based energy demand forecasting method by approximating the Shapley values leveraging DeepLIFT explainability technique. We present the experimental results on two different smart home energy consumption datasets and demonstrate that our method achieved state-of-the-art results in forecasting future energy demand compared to known related methods.

On top of that, our methods demonstrated the explanations in terms of the contribution of appliances and associated time. To evaluate the efficiency of the generated explanations, we introduced new evaluation metrics based on the monotonous relationship with the ground truth. We found that the generated explanations can capture the efficient contributions of different appliances. We also compared the findings with one of the relevant works in the same dataset.

Next, this part presents the findings on thermal comfort preference prediction for automatic heating and cooling systems. Chapter 7 first intro-

duced four new supervised feature selection techniques to filter redundant, noisy, and irrelevant features. We empirically investigated and observed that many features are correlated and redundant. Then, we come up with the feature selection techniques.

Then, we focus on the next challenge, data deficiency, because datasets of thermal comfort prediction lack enough data to train the ML model. We then introduce GAN to synthetic data samples to get rid of the challenges and train the models. Then, we evaluated the models, and in terms of all evaluation metrics, our method achieved better performance than existing methods. Finally, we present the global explainability of the models at the end.

5 Explaining AI Decisions: Towards Achieving Human-Centered Explainability in Smart Home Environments.

The content of this chapter has been presented in the 2nd World Conference of eXplainable Artificial Intelligence (xAI2024) which has been held in Malta in July 2024 and the paper has been published in the proceedings of the conference by Springer Nature. The information of the published paper is given as follows:

Article Information: Md Shajalal, Alexander Boden, and Gunnar Stevens, Delong Du, Dean-Robin Kern. 2024. Explaining AI Decisions: Towards Achieving Human-Centered Explainability in Smart Home Environments. In *Proceedings of the 2nd World Conference on eXplainable Artificial Intelligence 2024 (xAI2024)*. Communications in Computer and Information Science, Springer Nature Switzerland, 418–440. https://doi.org/10.1007/978-3-031-63803-9_23 (*Reproduced with permission from Springer Nature*)

Abstract

Smart home systems are gaining popularity as homeowners strive to enhance their living and working environments while minimizing energy consumption. However, the adoption of artificial intelligence (AI)-enabled decision-making models in smart home systems faces challenges due to the complexity and black-box nature of these systems, leading to concerns about explainability, trust, transparency, accountability, and fairness. The emerging field of explainable artificial intelligence (XAI) addresses these issues by providing explanations for the models' decisions and actions. While state-of-the-art XAI methods are beneficial for AI developers and practitioners, they may not be easily understood by general users, particularly household members. This paper advocates for human-centered XAI methods, emphasizing the importance of delivering readily comprehensible explanations to enhance user satisfaction and drive the adoption of smart home systems. We review state-of-the-art XAI methods and prior studies focusing on human-centered explanations for general users in the context of smart home applications. Through experiments on two smart home application scenarios, we demonstrate that explanations generated by prominent XAI techniques might not be effective in helping users understand and make decisions. We thus argue for the necessity of a human-centric approach in representing explanations in smart home systems and highlight relevant human-computer interaction (HCI) methodologies, including user studies, prototyping, technology probes analysis, and heuristic evaluation, that can be employed to generate and present human-centered explanations to users.

Keywords

Explainable AI (XAI), Human-Centered XAI, Demand Forecasting, Machine Learning, Smart Home

5.1 Introduction

Due to advancements in sensor technology and machine learning (ML) over the past decades, smart home applications can now provide residents with the ability to monitor and control connected appliances via sensors [278]. These applications can even make decisions automatically using ML-driven techniques rather than relying on simple timetable logic. In the smart home energy domain, one notable energy-aware smart home application might be appliance-level energy-demand forecasting to make users more aware and help them optimize their energy consumption practices [278, 162]. Adjusting the heating system to provide a comfortable and healthy household and work environment based on predicting individuals' thermal comfort preferences can be another fascinating energy-related smart home application [11, 283, 282]. Other applications also often utilize complex ML models to make decisions, such as human activity recognition within the home, identification of energy-intensive activities for different household tasks [302], fall detection and health monitoring [214], and energy optimization [163, 278].

AI-based applications in smart home systems are becoming increasingly popular as homeowners aim to enhance their living environment while reducing energy usage. Previous studies [161, 162, 163, 311, 278] have modeled energy demand forecasting in smart homes using AI techniques, including Deep Neural Networks (DNNs), Convolutional Neural Networks (CNNs), Auto-Encoders (AE), and Long-Short-Term Memory (LSTM). Classical Machine Learning (ML)-based predictive models have also garnered attention for predicting personal thermal comfort in

indoor environments, thereby enhancing resident comfort [283, 1, 64, 94, 106, 299, 258, 237]. The complexity and opacity of these ML models often hinder their adoption in real-world scenarios due to difficulties in aiding users' decision-making. Since ML models can be very complex, involving thousands to millions of model parameters (i.e., deep learning models), they are often referred to as *black-boxes*. Decisions from black-box models can be unintelligible and may surprise users with unexpected predictions. In such cases, users require explanations to comprehend the predictions. Recently, there has been significant interest in elucidating the decisions of ML models across various fields, including language processing [285], financial analytics [277], e-commerce [277], medicine and health [151], bioinformatics [150], and smart home applications [163, 150, 278]. This effort to clarify ML models' decisions is referred to as eXplainable Artificial Intelligence (XAI).

XAI aims to develop AI systems that can provide clear explanations about the decision-making processes and the predicted decisions [21]. The "black-box" issue can lead to users' mistrust and confusion about these technologies. To improve users' trust and understanding, *Human-centered* explanations[4] can be a game-changer by providing clear and effective explanations to users, enabling them to troubleshoot issues and customize devices to suit their needs [88]. However, many XAI methods have been introduced and developed to explain the models' decision-making procedures and the reasons behind specific predictions. Most of them are proposed to debug and improve models' performance [278, 87, 145]. In the context of smart home application scenarios, users in households are generally laypeople and may not have sufficient knowledge to understand technical explanations (i.e., even AI developers might struggle to understand explanations). Therefore, this paper advocates the need for easily understandable explanations for gen-

[4]Throughout the paper, "Human-centered XAI" and "user-centered XAI" are used interchangeably

eral users to make sense of predictions, which we refer to as "Human-centered" explanations.

While some studies have attempted to make models interpretable in the context of smart homes, most are focused on explaining predictions to improve performance and debug models [215, 112, 162]. As noted earlier, smart home users often lack the technical expertise necessary to understand many of the suggested explanations [247, 239, 145]. Additionally, research has shown that end-users have diverse perspectives when trying to make sense of smart home systems [55], and these perspectives can evolve over time. Various studies have demonstrated that current XAI techniques fail to produce human-centered explanations that assist general users in understanding the decision-making process and the reasons behind specific predictions [251, 38]. Given these findings, generating human-centered explanations for complex smart home applications is more challenging than might be initially assumed, especially considering the diverse backgrounds of general users.

Smart home systems encompass various energy consumption-related subtasks, including energy demand forecasting [278, 163], appliance-level consumption predictions [161], energy intensity identification for different household activities [302], and thermal comfort prediction [282, 283] for efficient heating systems. These systems are more complex than classical classification or regression tasks, and their collective outcomes contribute to the overall functionality of smart homes. However, the complexity of these systems can lead to decisions that are difficult for general users to understand, potentially hindering their adoption in real-world settings. This paper emphasizes the importance of human-centered explanations in smart home applications by reviewing the progress of state-of-the-art technical and human-centered XAI studies. We focus on two energy consumption-related application scenarios within smart home settings: energy demand forecasting and the

prediction of personal thermal comfort preferences for smart heating systems. We conduct experiments by applying deep neural network-based energy forecasting and ML models on benchmark datasets to model thermal comfort preferences. To uncover the complexity of the predictions, we apply current prominent XAI methods to identify facts by presenting the explanations in various forms.

We present explanations generated by multiple XAI methods and analyze their understandability. We then identify the associated challenges that must be considered when generating human-centered explanations. The need for user-friendly explanations in these contexts illustrates the challenges in understanding complex decisions. Finally, we highlight several HCI methodologies that could be beneficial in achieving human-centered XAI in smart home applications. The contributions of this paper are threefold:

- We present state-of-the-art XAI methods, discuss progress towards human-centered XAI and highlight research gaps that hinder their immediate application in smart home systems.
- Through careful analysis of experimental results based on prominent explainability methods in two sub-tasks, we argue for the need for human-centered explainability to understand decisions from complex AI-enabled smart home applications.
- We also emphasize several human-computer interaction (HCI) methodologies, including user studies, prototyping, technology probes analysis, and heuristic evaluation, to achieve human-centered XAI-enabled smart home applications.

The rest of the paper is organized as follows: In Section 5.2, we present state-of-the-art technical XAI methods and human-centered XAI studies. In Section 5.3, we present two smart home application scenarios that illustrate the need for human-centered explainability in smart

homes. We conducted experiments applying prominent ML techniques and XAI methods to predict and explain the models' decisions in Section 5.4. We also demonstrate why the provided explanations are insufficient for users to understand them. In Section 5.5, we highlight multiple HCI methodologies that demonstrate how to elicit requirements and design human-centered explanations. Finally, Section 5.6 concludes the paper by outlining future directions.

5.2 Human-Centered XAI and Current Progress

The broad goal of XAI is to enable general users to understand the working principles of AI models and their decisions through explanations. Several terms, such as "*interpretable AI*" and "*transparent AI*", are used interchangeably to describe the exact purpose of XAI [251]. The major objectives are similar: to make AI models and their decision-making processes understandable to general users through explanations. However, the number of methods focusing on a user-centered perspective is significantly lower than methods prioritizing improving model performance with technical explanations. As a result, the broad goal of XAI has not yet been fully achieved, which could potentially hinder the adoption of AI models in real-world applications.

This section provides a concise overview of the current cutting-edge approaches in explaining opaque decisions made by machine learning and deep learning-based predictive technologies, focusing on both technical and human-centered XAI. The primary aim of XAI advancements is to clarify the overall priorities and predictions of models for developers, facilitating the debugging process and enhancing model performance. While the field of technical XAI offers a wide range of techniques aimed at model improvement, it is important to note that the consideration of human-centered perspectives is relatively limited [251]. To present a comprehensive understanding of the literature, this section focuses

on three key issues: i) technical XAI, which provides a background of XAI; ii) human-centered XAI; iii) research gaps in the development of human-centered applications for smart home technology.

5.2.1 Technical XAI

The methods used in explainable artificial intelligence can be broadly classified into two categories: global and local explanation methods [209, 147]. Global explainability methods aim to identify the overall priorities of a predictive model and provide a summary of the decision-making process. In contrast, local explainability methods focus on understanding why a specific predictive decision was made, shedding light on the insights associated with that decision. Additionally, explainable AI approaches can be categorized as either model-agnostic or model-specific. Model-agnostic approaches can be applied to any predictive model to explain its predictions, while model-specific explainable techniques are designed for particular predictive algorithms [277].

One prominent and widely used XAI method is SHapley Additive Explanation (SHAP) [195], which generally explains the global priorities of models and highlights the most and least significant features according to their contribution. Following the game theory concept, SHAP computes each feature's weight that contributes to a specific decision. Moreover, SHAP can also provide local explanations for specific decisions. To explain deep neural network-based predictive models, well-known methods such as Grad-CAM [271], based on gradient localization, and Layer-wise Relevance Propagation (LRP) [211], which redistributes the output weight using backward propagation, can be used. These methods rely on saliency maps to explain decisions and can be applied to image and text-based applications. To uncover a DNN model's decisions, DeepLIFT (Deep Learning Important FeaTures) [293] has been introduced to identify important features for specific predictions of a model. Lakkaraju et

al. [172] proposed an XAI technique for AI experts to make the model and its behavior understandable by using subspace explanations. To provide explanations for experts, Schetinin et al. [263] introduced probabilistic interpretation for Bayesian decision tree models.

Another way to explain models' predictions to users is through example-based explanations. With these types of explanations, XAI approaches provide similar instances that match the corresponding samples with the same decision. Several example-based interpretability models that consider similar prototypes and criticisms have been introduced to help users understand why a certain decision was made [156, 117]. To explain predictions for any complex deep learning model, Local Interpretable Model-Agnostic Explanations (LIME) [244] can create a *surrogate model* that mimics the performance of the complex model. The *surrogate model* is explainable, and its performance is quite similar to that of the original model. With LIME, any particular prediction can be explained for tabular and textual data by highlighting positive and negative features corresponding to the predicted decision.

To comprehend specific events, humans sometimes look for explanations that involve significant changes in the attributes of a sample, which can overturn the original decision. These are known as counterfactual explanations. They help users understand "*why a different prediction was not possible?*" or "what changes could modify the final prediction?" Counterfactual explanations can also be illustrated by introducing new example samples that could reverse the decision, known as "what-if" scenarios. Various XAI approaches have been developed to explain "what-if" and counterfactual scenarios [310, 213, 315]. Other methods, such as the partial dependence plot (PDP), individual conditional expectation (ICE), and DiCE [213], are also used for similar purposes. To integrate multiple types of explanations into a unified framework, several open-source toolkits are available, including captum [166], AIX360 [24], and

Anchor [246].

5.2.2 Human-centered XAI

The progress in the field of XAI, thus far, in generating technical explanations to interpret models for AI practitioners for debugging and improvement, far exceeds the advances made in human-centered perspectives of XAI. However, a significant number of scientific studies have emerged that focus on the user-centric perspective. In this section, we review some notable works on the evaluation of human-centered XAI.

Bell et al. [38] demonstrated that the model-agnostic explanations provided by one of the most prominent XAI techniques, SHAP, are not sufficiently comprehensible for general users. They assessed the effectiveness of these explanations through an empirical study involving non-technical participants in two distinct application areas, education, and finance [38]. Similarly, Abdul et al. [2] explored the trade-off between accuracy and simplicity in explanation presentation and proposed a cognitive generalized additive model (COGAM) for human-centered explanation delivery. A generalized XAI design principle was introduced to facilitate the presentation of local explanations to non-expert users by contextualizing the exploration of feature importance. An empirical study involving more than 80 participants was conducted to evaluate the effectiveness and user satisfaction with the explanations provided.

Chromik et al. [67] investigated the capability of non-expert users to understand and form a mental model of global explanations to comprehend the behavior of a model. Their findings indicated that global explanations are insufficient for understanding the overall model behavior. In a separate study, Hase et al. [122] to measured the *simulatability* of various XAI techniques, including LIME, Anchor, prototypes, and decision boundaries, for tabular and textual data. They concluded that

LIME achieved good simulatability in a few cases, and prototype models were useful for counterfactual explanations. Another study on simulatability found that current interpretability methods are inadequate in explaining model behavior [189, 251]. However, multiple studies assessing the effectiveness of incorporating XAI approaches in real-world decision-making concluded that the evidence does not align with the goals of XAI [188, 30, 235, 317].

5.2.3 The Research Gap

The related works discussed above, addressing both technical and human-centered XAI, reveal a significant research gap in designing XAI systems for general users. Although current XAI systems can provide interpretable explanations in various forms—including global, local, counterfactual, and example-based explanations—these methods often fail to make predictions and model behavior understandable to non-expert users. Yet, one of the primary goals of XAI is to enable lay users to comprehend the predictions. Furthermore, applications in smart homes, such as energy demand forecasting, consumption practices, smart heating systems for comfortable home environments, and thermal comfort preferences with AI-enabled prediction systems, necessitate diverse perspectives for implementing user-centered XAI adoption.

Consider the complex task of developing a human-centered XAI energy prediction system for a smart home. This task is challenging due to the system's internal complexity and the diverse backgrounds of non-expert users. While significant progress has been made in technical XAI for applications related to classification and regression, the complexity of energy forecasting, which involves two different dimensions—time and characteristics or features—makes it difficult to represent the underlying factors behind the forecast. Moreover, the explanations must be designed and presented in a human-centered manner, which poses ad-

ditional challenges for creating human-centered explanations in smart home applications.

Another smart home application involves the automatic control of the heating system, which can be based on a prediction model of the inhabitants' thermal preferences. This system generally utilizes physiological and environmental data from the inhabitants, as well as weather data, to predict their thermal comfort preferences [283, 282]. Based on these predictions, the smart home application then adjusts the heating system to control the indoor temperature. Since this system utilizes data from various sources, it is crucial for users to understand how and in what context their data are used. Therefore, we require human-centered, easily understandable explanations from such systems to ensure transparency and trust.

5.3 Human-Centered Explainability in Smart Home

Smart homes equipped with AI-driven applications, including energy demand forecasting, consumption routines analysis, heating systems monitoring, and controlling indoor temperature based on occupants' thermal preference prediction systems, are becoming increasingly popular [278, 140]. These systems often use complex ML-based prediction systems that should be understandable to the household inhabitants [282, 278, 163]. The adoption of advanced XAI within complex smart home systems, offering human-centered explanations, could be transformative. As a result, non-expert smart home users would gain a clear and concise understanding of how their systems predict consumption and preferences, analyze their data for daily consumption practices, and control the indoor environment to ensure a comfortable living space. Incorporating current XAI techniques and including users in the development loop to consider their perspectives could significantly enhance the provision of human-centered explanations.

We present an overview of human-centered XAI in smart home applications with occupants actively involved, as shown in Fig. 5. The prediction system should consider the occupants' preferences and intentions to provide meaningful explanations. Initially, ML-enabled predictive models learn from preprocessed data to make future predictions and forecasts. When implementing XAI models to elucidate the decision-making processes of these models and offer explanations for individual predictions, it is crucial to incorporate a human-centered perspective. In this context, human-computer interaction methods are essential for analyzing user feedback. Consequently, effectively presenting these explanations can enhance the adoption of AI-enabled systems in real-world smart home settings. To highlight the importance of human-centered explainability methods in smart home applications, we selected two relevant and extensively investigated problems within the smart home context.

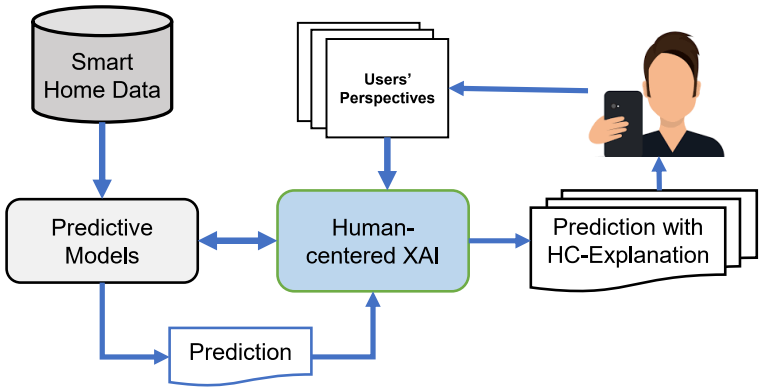

Figure 5: An overview of a human-centered XAI-enabled Smart Home systems

To select the problem domain in the context of smart homes, we focus on applications related to household energy consumption. This emphasis allows us to explore areas that enhance resident comfort while also optimizing energy use. Additionally, smart applications should offer in-

sights to help people optimize and reduce their energy costs. Therefore, we have chosen two applications: household energy demand forecasting and automatic control of heating based on predictions of thermal comfort preferences. Both applications are concerned with energy consumption practices. The first application is an energy demand forecasting system designed to increase awareness and optimize household electricity use. The second involves modeling occupants' thermal comfort preferences to ensure a comfortable and healthy indoor environment. In the remainder of this section, we present two relevant smart home sub-tasks that require human-centered explanations for successful adoption.

5.3.1 Household energy demand forecasting

Energy consumption in residential and commercial buildings significantly exceeds that in other sectors [282, 278]. Additionally, the price of energy is continuously increasing worldwide. However, this heightened consumption also results in significant CO2 emissions, posing a serious threat to global warming and the environment [282]. Smart home systems address these concerns by providing future energy demand forecasts for households based on historical energy usage data collected from various appliances, thanks to advanced sensor technology [163]. Such predictions of total energy demand for the upcoming month or week might raise household members' awareness of their energy-related activities, potentially encouraging them to optimize their energy usage.

Nevertheless, energy demand forecasting systems typically rely on highly sophisticated ML and DL techniques [161, 163, 87], which perform complex calculations and lead to opaque decision-making processes. Consequently, some predictions may surprise users with unexpected outcomes. For example, if a forecasting system predicts a high (or low) total energy consumption for washing machines in the next month, users might be taken aback, as they might perceive washing machines to con-

sume less (or more) energy than predicted. In such scenarios, users require easily understandable explanations from the systems to comprehend and build trust in AI-enabled systems, thereby enhancing their adoption.

Unlike other classification and regression tasks, providing explanations for energy demand forecasting systems is non-trivial, as it involves multivariate time series forecasting [278]. The explanations in such systems cover two dimensions: they relate to features or attributes and time. Consequently, explainability methods must capture the impact of time and features, making understanding explanations for time series forecasting more challenging for users. Research into human-centered explainability methods in this field is crucial to address this issue. This research will enable inhabitants to understand why a certain energy demand is anticipated for the upcoming month and will facilitate the development of more optimal energy consumption plans based on factual explanations.

5.3.2 Occupants' thermal comfort preference modeling

Indoor thermal comfort is essential for the well-being, comfort, and work productivity of inhabitants [282, 283]. With recent advancements in efficient sensors and smart home appliances, AI-driven heating, ventilation, and air conditioning (HVAC) systems can monitor and control the indoor environment. These systems have gained considerable attention for applying machine learning techniques to automatically control parameters related to the comfort of the indoor environment [291, 283]. Personal thermal comfort preferences vary widely from person to person, making the prediction of individual preferences crucial for providing occupant-level comfort in households [282, 283]. Based on these preference predictions, the heating system can be automatically adjusted to control the temperature at the occupant's location.

Householders often struggle with inconsistent temperatures in their homes, especially during extreme weather conditions, leading to discomfort and high energy costs [312]. Maintaining a consistent temperature throughout the house can be challenging, causing the HVAC system to work harder to maintain a comfortable temperature using AI-enabled computational models. However, these complex AI systems often lack interpretability. Traditional temperature control systems typically react by adjusting the temperature only after it has already started to fluctuate, resulting in uncomfortable temperature swings and inefficiencies. To address this problem, a more proactive approach to temperature control is needed. Various methods have been introduced to predict personal thermal comfort preferences using complex machine learning and deep learning models [1, 92, 299, 282]. These automated systems, powered by complex AI models, can monitor and control the indoor environment. However, the decision-making process and the rationale behind specific predictions and actions often remain unclear to the inhabitants, including AI practitioners themselves. As XAI methods progress, it is crucial to make the explanations human-centered, enabling household occupants to understand the reasons behind predictions and the decision-making of the models. This would facilitate the adoption of such complex models and ensure successful smart home applications.

5.4 Experiments and Analysis

This section presents the details of the experiments we conducted on the two aforementioned smart home scenarios. We carried out experiments by training various predictive models on two distinct datasets collected for both applications. Initially, we trained predictive models and subsequently applied two well-known XAI methods, namely SHAP [195] and DeepLIFT [293]. We then analyzed the generated explanations for both smart home applications and sought to identify reasons why these ex-

planations might not be sufficient for smart home users to comprehend. This section presents the experimental results, generated explanations, and their analysis for each scenario.

5.4.1 Energy demand forecasting in smart home

Experimental Settings: We conducted experiments on the REFIT dataset [217], which includes energy consumption data collected from 20 diverse households. The data encompasses various appliances such as the Fridge-Freezer, Tumble Dryer, Washing Machine, Dishwasher, Desktop Computer, Television, Microwave, Kettle, and Toaster, with energy consumption recorded at 8-second intervals. We modeled the weekly energy demand forecasting problem using a classical LSTM-based model. The LSTM-based forecasting model features 10 total features, a sequence length of 7, two hidden layers, 64 hidden units in each layer, 100 epochs, a learning rate of 0.001, and a batch size of 64.

Results: The performance of the LSTM-based forecasting model to predict upcoming weekly energy demand is presented in Table 1. We can see that the performance is quite effective in terms of four different evaluation metrics, including mean squared error (MSE), Root-MSE (RMSE), mean average error (MAE), and Mean absolute percentage error (MAPE). The performance across different households varied widely. For house 5, the forecasting performance is better than that of another household in terms of MAPE and MSE. On the other hand, for house 13, LSTM achieved the best performance in terms of MAE and RMSE. However, presenting the forecasting performance here makes sense in that the generated explanations for decisions can better capture facts and reasons.

Explainability: To explain the predictions from the model, we uti-

Table 1: Prediction performance in forecasting energy demand on 4 different households of REFIT dataset

House	MAE	MAPE	MSE	RMSE
House 2	0.1351	0.5421	0.03	0.1752
House 5	0.0768	0.4075	0.01	0.10
House 8	0.1934	0.4618	0.0451	0.2095
House 13	0.0435	0.8175	0.003	0.0564

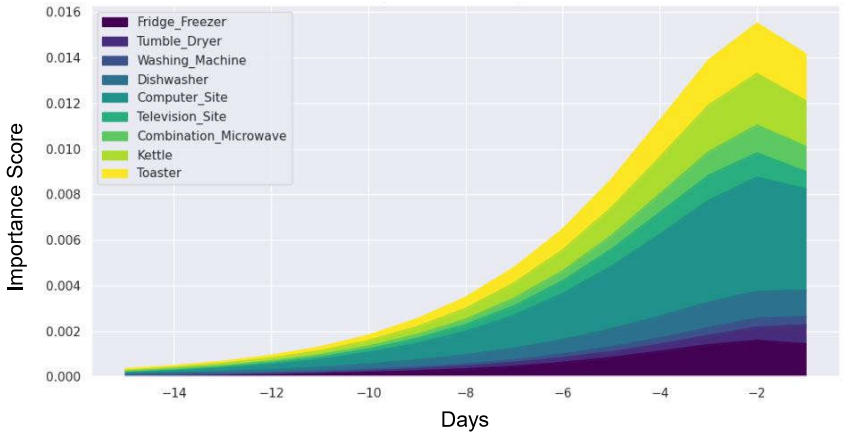

Figure 6: Explanations for weekly energy demand forecasting highlighting the contributions of different appliances.

lized Deep Learning Important Feature (DeepLIFT), which approximates Shapley values to provide an explanation. This method combines the contributions of different appliances towards the overall prediction, thereby informing users about the activities responsible for the total energy consumption. The explanations for weekly energy forecasting, presented in Figure 6, show how the contributions of different appliances vary over time. These explanations are fairly technical, indicating that the contributions of different appliances change with time. Previous studies have shown that general users often struggle to understand even straightforward explanations generated for binary classification tasks.

On the other hand, energy demand forecasting is an even more challenging task, and the explanations provided differ significantly. The dimension of these explanations also relates to time. As a result, there is

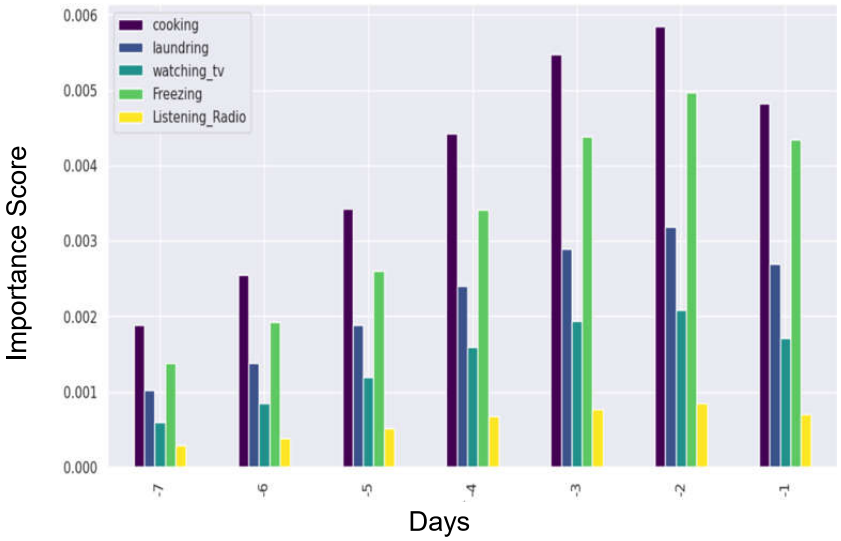

Figure 7: Explanations for weekly energy demand forecasting highlighting consumption activity corresponding to the time (day)

a need for explanations that help laypeople understand how these AI-enabled forecasting systems make decisions in their homes. To simplify, we sum up the contributions of different appliances and present the explanation using a bar chart. Figure 7 shows the contributions of different household activities responsible for overall energy consumption, with cooking activities expected to consume the most energy. What use are these explanations to a user? While they provide some insight into the activities most responsible for energy use, they do not enable users to optimize their energy consumption practices with the current form of explanations. It is also clear that understanding explanations generated by a single successful XAI technique can be challenging. The impact of different activities on overall consumption varies across different days. Non-expert users may develop a negative perception when confronted with such complex explanations. Therefore, explanations should be easy to follow, trustworthy, and tailored to a human-centered perspective to facilitate sense-making.

Table 2: The performance of different classical machine learning models on predicting personal thermal comfort preference. The best results are in **bold**.

Model	Kappa	Accuracy	AUC
Decision Tree	0.6609	0.8315	0.8810
Support Vector Machine	0.5470	0.8167	0.9198
K-nearest Neighbor	0.3810	0.7589	0.8101
Gaussian Naive Bayes	0.4311	0.7148	0.7689
XGBoost	**0.6774**	**0.8457**	**0.9487**
Random Forest	0.5901	0.8195	0.9258

5.4.2 Personal thermal comfort preference prediction

Experimental Settings: For the second smart home sub-task, "*thermal comfort preference modeling*," we conducted experiments using a wearable dataset collected by UC Berkeley. The dataset originates from a field experiment involving 14 subjects living in Berkeley and San Francisco [283]. It contains a total of 3848 samples from all participants, categorized into their thermal comfort preferences: "Cooler" (class 0), "No Change" (class 1), and "Warmer" (class 2). Further details about the dataset can be found in [283]. After preprocessing the values of the features, we applied a feature selection technique to identify relevant features, resulting in the selection of 32 different features [283]. We then trained six different prominent classical machine learning models, including Decision Tree (DT), Support Vector Machine (SVM), K-nearest Neighbor (KNN), Gaussian Naive Bayes (GNB), XGBoost (XGB), and Random Forest (RF).

Results: The performance of all trained machine learning models is presented in Table 2. We evaluated their performance using metrics such as Cohen's Kappa, Accuracy, and Area Under the Curve (AUC). Given that the dataset was quite imbalanced in terms of the number of samples for different classes, we employed metrics that can effectively evaluate the performance of ML classifiers on imbalanced data. For this purpose, we applied Cohen's Kappa and AUC. From Table 2, it is evident

that XGBoost achieved better performance across all evaluation metrics. The performance of the other five models was also quite consistent compared to the baseline models on the same dataset by [283].

Explainability: We incorporated the prominent XAI method SHAP to generate explanations. We illustrated the global and local explanations in Figure 8 and 9. Figure 8 shows the global explanations, indicating that thermal sensitivity is the most important feature for modeling personal thermal comfort. The next most important features are cold experience, age, weight, and work hours. As an AI practitioner, one can understand which features the model prioritizes for the overall decision. However, as an inhabitant of the household, some might struggle to derive meaningful insights from this explanation.

In Fig. 9, we illustrate the explanations for a particular subject's thermal preference at a certain time in terms of a waterfall plot; these explanations follow the global explanations in a broader sense. We can observe that the most important features are thermal sensitivity, height, and temperature in the ankle. Once again, developers can leverage these insights to enhance the models' performance by canonicalizing and modifying them. Nevertheless, general occupants may find it challenging to derive actionable insights from this information.

5.5 Challenges and HCI Techniques for Human-centered Explainability

After carefully analyzing the current literature on XAI and human-centered XAI (Section 5.2) and generating explanations using two established XAI methods across two different smart home scenarios, we have identified challenges that need addressing to achieve human-centered XAI in these applications. Smart home applications span various prob-

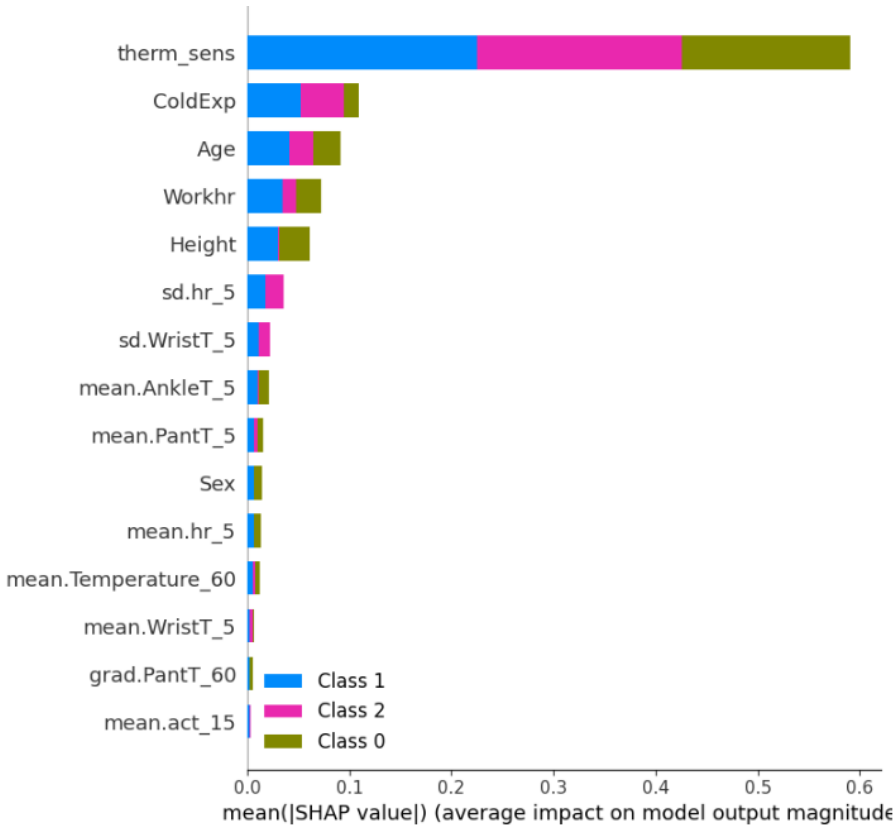

Figure 8: Global explanation for personal thermal comfort preference prediction highlighting model's overall priorities.

lem domains, including classification (e.g., PTC preference prediction), time-series forecasting (e.g., energy demand forecasting, energy billing), and regression (e.g., activity recognition). Thus, the challenges for achieving user-centric explainability must be tailored to each specific domain. While technical explanations are useful for debugging and optimizing model performance, they do not typically aid general users' decision-making. If explanations do not meet users' needs, this can result in a lack of trust in the model and resistance to its use [21]. Human-centered XAI tools, designed with end-users in mind and providing contextually relevant and understandable explanations, can overcome these challenges [88]. To offer human-centered explanations for smart home systems, it is crucial to employ user-focused XAI tools, provide understandable explanations, and address the users' inquiries and

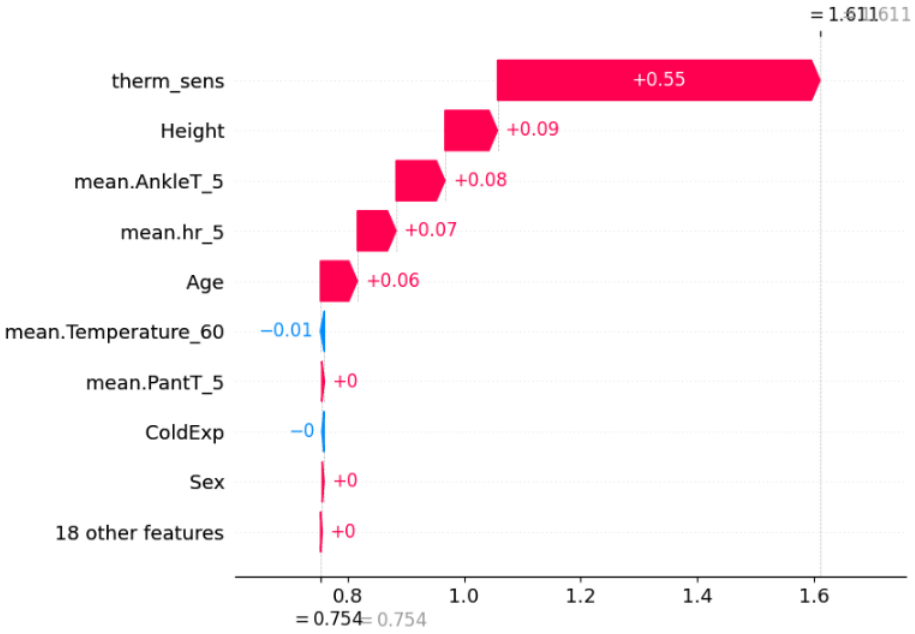

Figure 9: Explanation for a decision that predicts the occupants felling "*warmer*".

concerns [88]. Addressing the following challenges will help in generating human-centered, XAI-enabled explanations, ultimately improving the adoption of AI systems in real-world applications.

5.5.1 Challenges in achieving human-centered explainability

When discussing the challenges of human-centered explainability, we adopt the notions of syntax, semantics, and pragmatics, which are traditionally utilized in linguistics but are effectively applicable in identifying challenges in smart home explanations. These concepts aid in understanding and improving how users design and perceive explanations.

The syntax level of explanations refers to the presentation and visual encoding of explanations, emphasizing user-friendly choices in colors, fonts, layouts, and chart design to minimize cognitive load and simplify interactions with smart devices [300]. Although charts generated by XAI frameworks such as SHAP or LIME are highly accurate, they tend

to be mathematical and unengaging. In technical domains, this might not pose a problem; however, for lay smart home users, factors such as aesthetics, playfulness, comprehensibility, appropriate measurement units, and careful wording are critical [56]. For instance, Schwartz et al. [269] demonstrate that technical units like *kWh* (kilowatt-hour) or *kg CO2 eq* (equivalent to the effect of one kg of CO2 emission) are too complex, whereas laypersons typically prefer money as a well-known, easy-to-interpret unit. Another challenge in the visual design of explanations is the small screen size [300], as most people interact with their smart home through smartphones or wall-mounted interfaces, necessitating the simplification of complex explanations for small-screen visualization.

The semantic level of explanations concerns how explanations are interpreted and what mental models are generated [300]. In the context of smart homes, for instance, many individuals possess incorrect mental models of heating systems, leading to improper heating behaviors [154]. Therefore, explanations must be mathematically precise and assist users in developing accurate mental models. In this regard, Schwartz et al. [269] demonstrate that people often rely on ethno- or folk-methods to construct their mental models. Regarding domestic energy consumption, people interpret the information provided by eco-feedback systems using money as the preferred unit to assess appliance consumption, relate consumption to their habits, or compare their consumption with others' [269]. Explanations should leverage these folk methods to help people build accurate mental models [154, 269].

The pragmatic level of explanations refers to the context-dependent, practical significance of explanations as they apply to the user's daily life. For example, explanations can serve various purposes, such as building trust in smart home systems [136], increasing the energy literacy of residents [267], supporting reflection on wasteful consumption

habits [269], and prompting actions to detect and replace inefficient appliances. Additionally, they enhance predictability and accountability [136] and support the co-performance of controlling domestic appliances [168].

Explanations should be tailored to these pragmatic considerations to be effective. For example, simple recommendations accompanied by "what-if" explanations [315] are more effective for actions such as detecting and replacing wasteful appliances. In contrast, more elaborated, cause-effect-oriented explanation approaches [125] are more helpful in enhancing energy literacy. By considering all three aspects—syntax, semantics, and pragmatics—we can better identify and overcome the specific challenges associated with explaining smart home technologies, making them more user-friendly and aligned with human-centered design principles.

In the following, we discuss these challenges across three scenarios.

1. **Making predictions and autonomous actions interpretable for users:** In the context of energy demand forecasting, XAI tools can provide understandable explanations for the predictions made by the forecasting model, highlighting the most influential factors contributing to the prediction. These explanations should be presented in a user-friendly format that aligns with the user's intuition and understanding of the problem. Visualizations such as graphs and charts can be used to illustrate the data and make it easier for users to comprehend the predictions. In the case of an autonomous action performed by an AI-enabled system (e.g., adjusting the room temperature based on predicted thermal comfort preferences), the system should provide clear explanations for why that particular action was taken.

2. **Providing insights for an optimal energy consumption plan:**

Human-centered, XAI-powered smart home applications should offer users insights into their optimal energy consumption plan by considering their energy usage patterns, environmental conditions, and preferences. The system might involve developing a personalized energy optimization plan, such as adjusting the temperature based on the user's daily routines or turning off lights in unoccupied rooms. The human-centered XAI-enabled system should provide suggestions and explanations for energy-saving practices and offer feedback on the impact of these practices on energy consumption.

3. **Making users aware of energy consumption:** XAI tools should help users understand how their energy consumption patterns affect the home environment and their energy bills. This can be achieved by providing real-time feedback on energy usage, highlighting areas where energy could be conserved, and suggesting energy-efficient practices. For instance, the XAI tool can send alerts or notifications to remind users to turn off lights or appliances when not in use.

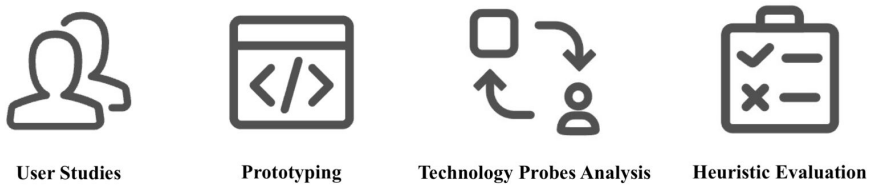

Figure 10: HCI techniques to enhance human-centered explainability

5.5.2 HCI techniques to enhance human-centered explainability

Focusing on human-centered XAI can align with users' needs and preferences, increasing their trust and understanding of smart home AI systems [88]. This alignment is particularly significant as we strive for

a more energy-efficient and comfortable living environment with smart home systems. Thus, Incorporating HCI methodologies becomes effective in advancing smart home, human-centered XAI implementation. The following non-linear process in Fig. 10 illustrates this approach: User Studies [248], Prototyping [222], Technology Probes Analysis [132], and Heuristic Evaluation [221].

1. **User Studies:** User studies employ detailed observation, daily tracking, and interviews based on praxeological grounded design methods to understand contextual user practices in smart home environments [248]. This approach aims to gather deep insights into how users interact with XAI artifacts of smart home technology, facilitating a refined understanding of social practices. By analyzing these interactions, developers can align XAI systems more closely with user habits and expectations. Such alignment increases user trust and enhances system transparency. These insights are crucial for tailoring AI functionalities to manage home heating or cooling systems efficiently. For example, based on collected data, AI could predict and adjust indoor temperatures to optimal levels just before users return home or during specific weather conditions, thus maximizing comfort without the need for manual adjustments [278, 283, 282]. These adaptive adjustments contribute significantly to improving energy efficiency and overall user comfort.

2. **Prototyping:** Developing several low-fidelity prototypes offers a cost-effective way to explore different designs and continuously gather user feedback. This process is characterized by iterative testing and refinement cycles, heavily using participatory design principles to align with user expectations and improve usability [222]. Through prototyping, designers can quickly adapt and evolve XAI features based on real user feedback, ensuring the sys-

tem is both intuitive and directly useful to end-users. Such an approach allows developers to fine-tune how XAI communicates energy-saving decisions to users. For example, the explanations might suggest reducing electricity use during off-peak hours as a cost-saving measure. By clearly explaining the rationale behind such recommendations, prototyping enhances the usability of XAI, making it easier for users to trust and follow the AI's guidance. This process not only aids in reducing energy costs but also helps in educating users about efficient energy practices.

3. **Technology Probes Analysis:** Technology probes are invaluable for understanding how users interact with and respond to XAI systems within smart homes, especially focusing on aspects such as interpretability, responsibility, and relevance of AI explanations. By deploying technology probes, such as smart meters that track energy usage and provide feedback and advice based on AI analysis, developers can gather rich data on user behavior and preferences [132]. These probes reveal user reactions to automated suggestions for optimizing energy consumption. The insights gained from technology probes allow developers to refine human-centered XAI explanations, ensuring they are both meaningful and actionable. This process enhances the system's usability and boosts users' understanding of and trust in the AI explanations, leading to more effective and sustainable smart energy management.

4. **Heuristic Evaluation:** Heuristic evaluation involves collaboration with subject matter experts, including HVAC, electrical, and social engineers, to assess AI systems' responsibility and user-centric design. This evaluation focuses on key aspects such as transparency, user control, and ethical considerations. Experts examine whether an AI system's explanations for recommending specific energy-saving measures are understandable and align with user

values [221]. Such evaluations are crucial because they help ensure that AI systems are designed with a strong emphasis on user-centric principles, which promotes a better understanding of and trust in the technology. Heuristic evaluation is particularly effective at identifying aspects of AI explanations that may not be evident through end-user testing alone. Doing so aids users in making informed decisions about their energy use, like understanding why certain settings are recommended for maximizing thermal comfort without excessive energy use [278, 283, 282].

5.6 Conclusions and Future Directions

Our research strives to achieve human-centered XAI for smart home applications, aiming to make complex AI-driven models understandable to laypeople. Through experiments on two smart home sub-tasks, we demonstrated the challenges associated with understanding decisions from such applications. We argue that current explanation generation techniques are insufficient for making general users comprehend these decisions. By identifying major challenges, we highlight the need for human-centered explainability and discuss how these challenges can be addressed using various human-computer interaction (HCI) techniques, including user feedback, co-performance considerations, and expert-user co-design. Future human-centered XAI can apply HCI techniques to understand users' requirements and preferences better, enabling them to understand decisions from AI-driven systems. Based on the outcomes of these techniques, we aim to develop human-centered explanations that facilitate user understanding and action.

However, challenges persist in generating human-centered explanations that foster trust and interpretability in AI systems. Future research directions include:

- Exploring natural language explanations and interactive interfaces
- Developing standardized frameworks to evaluate the general user experience of human-centered explainability
- Integrating HCI methodologies into the XAI development life-cycle to ensure user-centered and effective explanations

Future research on emerging user interfaces, including ubiquitous and pervasive technologies, can also advance the current state of the art on human-centered XAI in the context of the smart home energy domain. As we envision the integration of smart devices such as voice assistants, smart watches, and embedded sensors into everyday environments, these technologies provide a rich platform for deploying human-centered solutions to benefit sustainability. Such interfaces can offer intuitive and context-aware interactions, making AI explanations part of the natural user environment and activities.

Acknowledgment

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 955422.

6 *ForecastExplainer*: Explainable household energy demand forecasting by approximating shapley values using DeepLIFT

The content of this chapter has been presented as a poster in the Thirteenth ACM International Conference on Future Energy Systems 2022 (ACM e-Energy'22) and published as a short paper in the proceedings of the conference by ACM. The extended version has been published as a full length research article in *Technological Forecasting and Social Change* journal by Elsevier. The information of both papers is given as follows:

Information for Article 1: Md Shajalal, Alexander Boden, and Gunnar Stevens. 2022. Towards User-centered Explainable Energy Demand Forecasting Systems. In *Proceedings of the Thirteenth ACM International Conference on Future Energy Systems, (e-Energy '22)*. Association for Computing Machinery, New York, NY, USA, 446 – 447. https://doi.org/10.1145/3538637.3538877

Information for Article 2: Md Shajalal, Alexander Boden, Gunnar Stevens. 2024. ForecastExplainer: Explainable household energy demand forecasting by approximating shapley values using DeepLIFT. *Technological Forecasting and Social Change*, Elsevier. Volume 206, 2024, 16 pages. https://doi.org/10.1016/j.techfore.2024.123588

Abstract

The rapid progress in sensor technology has empowered smart home systems to efficiently monitor and control household appliances. AI-enabled smart home systems can forecast household future energy demand so that the occupants can revise their energy consumption plan and be aware of optimal energy consumption practices. However, deep learning (DL)-based demand forecasting models are complex and decisions from such black-box models are often considered opaque. Recently, eXplainable Artificial Intelligence (XAI) has garnered substantial attention in explaining decisions of complex DL models. The primary objective is to enhance the acceptance, trust, and transparency of AI models by offering explanations about provided decisions. We propose *ForecastExplainer*, an explainable deep energy demand forecasting framework that leverages Deep Learning Important Features (DeepLIFT) to approximate Shapley values to map the contribution of different appliances and features with time. The generated explanations can shed light to explain the prediction highlighting the impact of energy consumption attributes corresponding to time, such as responsible appliances, consumption by household areas and activities, and seasonal effects. Experiments on household datasets demonstrated the effectiveness of our method in accurate forecasting. We designed a new metric to evaluate the effectiveness of the generated explanations and the experiment results indicate the comprehensibility of the explanations. These insights might empower users to optimize energy consumption practices, fostering AI adoption in smart applications.

Keywords

Explainable energy demand forecasting, DeepLIFT, Shapley additive explanation, Deep learning, Human-centered explanation

6.1 Introduction

Smart home systems offer users the ability to remotely control and access household electrical appliances and monitor the environment through various sensors [140, 278]. Additionally, these systems can autonomously make decisions, such as adjusting the state of connected actuators (e.g., controlling the heating system to adapt room temperature) to enhance the dwellers' comfort [140, 76, 197]. While many smart applications follow simple timetable logic and classic automation paradigms, an increasing number of decisions are made using a combination of machine learning models [13, 181].

Energy demand forecasting using machine learning (ML) models has recently garnered significant attention in the literature. The objective is to make smart home users more aware of their future energy consumption [162, 341, 197, 161, 163]. These systems can even forecast the energy consumption for individual appliances [120], which enhances household members' awareness and encourages optimal electricity consumption practices. Since increased energy consumption can lead to higher household costs, people are expected to become more cautious and may modify their consumption behavior to decrease energy usage. How- ever, implementing such technology in the real world poses challenges due to its lack of transparency and trust. Users may not fully understand the reasons behind certain predictions and require more trustworthy explanations regarding the facts behind predicted decisions/recommendations. Alongside accurate energy demand predictions, one sensible approach to building trust and increasing transparency and fairness is to explain the predictions by highlighting important factors and time duration.

However, the underlying forecasting models are often black boxes for the end-users (even for AI practitioners), who don't have a clear understanding of the decision-making procedures of these prediction systems. As

a consequence, users might want factual explanations for why a particular decision has been taken on their behalf by the system [87]. They might have queries such as "*Why do we need this amount (more/less) of energy in the upcoming week/month*?". The plausible reason behind such an impression is that the system might surprise the user with an unexpected prediction (i.e., forecasting more/less amount of energy for next week/month compared to their expectation). However, providing explanations by highlighting the significant factors corresponding to the time duration might make them understand the reasons behind such predictions. Moreover, to plan the optimal energy consumption in the household, it would be more effective if they know which appliances might be responsible and consume more for the future overall predicted consumption. The relevant question can be "*How can I further optimize my energy consumption?*".

Providing comprehensive explanations from the system to address these questions would likely enhance the trust and transparency of AI models for end-users [277, 109]. Additionally, in accordance with the General Data Protection Regulation (GDPR), citizens of the EU have a civil right to be informed about how AI-based models that pertain to them make decisions [83]. The incorporation of explainability to ensure transparency, offering comprehensive explanations employing clear and interpretable facets unveiling DL-based fore- casting methods is expected. Explainable forecasting systems have the potential to augment seamlessly the overarching goals centered on technological advancement in AI and comprehending their corresponding societal ramifications.

The research field that focuses on explaining (and/or interpreting) the decision-making process is commonly referred to as eXplainable Artificial Intelligence (XAI). In recent years, there has been a significant interest in interpreting and explaining complex machine learning models [244, 245, 195, 72, 71, 121], enabling AI practitioners

and developers to enhance models' performance. The application of XAI has also received considerable attention in various fields, including bio-informatics [150], healthcare [5], finance [331, 108, 86], inventory management [277], natural language processing [149, 284] and so on [160, 318, 157]. However, there remains a need for further research on how to generate *human-centered* explanations that are accessible to end-users with no expert knowledge in AI theory and development [158].

As the human-centered design is highly context-specific, such research would arguably need to take into account the specific user needs of different domains, i.e., studying how explanations can be made meaningful to users in a specific pragmatic context and situated action. In our study, we focus on the domain of smart home technology, where such challenges are prominent but have been hardly studied to our knowledge. In addition, explaining the multivariate time series forecasting model is difficult [72], because the explanations might be two-dimensional, including both time and features. Hence, generating human-centered explanations for energy demand forecasting is a challenging task, especially when those are to be understood by end-users.

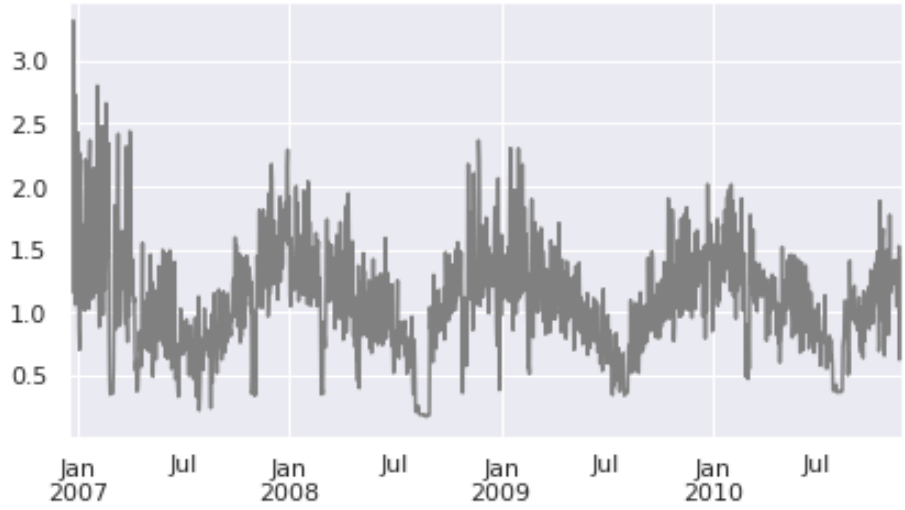

Figure 11: Daily distribution of global active power

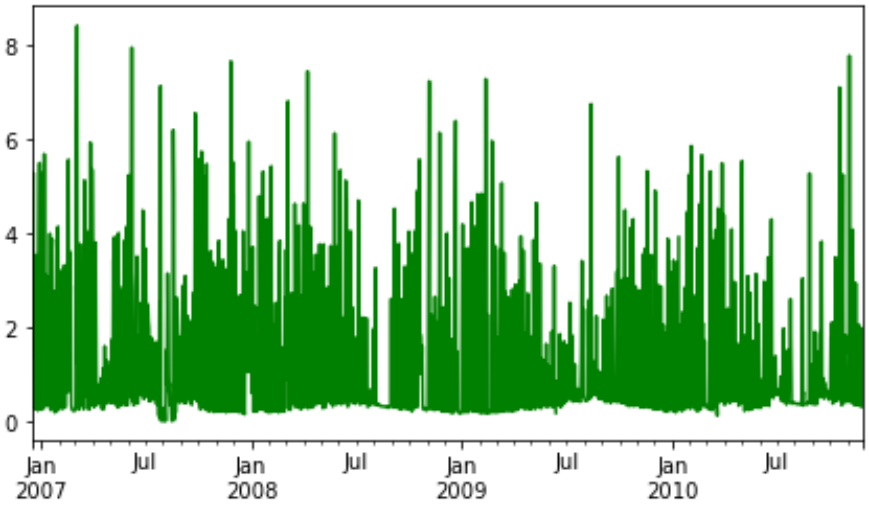

Figure 12: Daily energy consumption in the Kitchen

Let us discuss how different and complex the household energy demand forecasting problems are. Here, we discuss the problem in two different directions. Fig. 11 and 12 indicate the daily total energy consumed by the whole household area and kitchen, respectively. We can see that the energy consumption patterns are quite different. These complex and dissimilar consumption patterns also exist in some other places of the household area including the living room and laundry room. The irregularity in energy consumption for different appliances makes the forecasting problem challenging for any ML models. Moreover, the seasonal consumption patterns for the different household areas are widely varied over time. Therefore, forecasting household energy demand for different areas of the house is challenging.

In addition, time series forecasting models are dependent not only on the values of the features like classical classification or regression tasks but also dependent on time. Since the multivariate forecasting models are dependent on both the features and time, explaining specific predictions by highlighting important factors and corresponding time frames is very challenging. Moreover, unlike classification tasks, the progress of developing XAI tools for understandable explanation is much lower

for multivariate forecasting tasks. Therefore, explaining the decision for the time series forecasting model is a more formidable task than other classical classification or regression models.

In this paper, we present the underlying challenges to generate and present explanations for a particular prediction of an energy demand forecasting system. To explain the prediction of the energy demand forecasting system, we propose an explainable framework, *ForecasExplainer* by approximating Shapley values leveraging DeepLIFT to explain predictions made by the deep learning-based energy demand forecasting model. First, we developed an energy demand forecasting model applying long short-term memory (LSTM) networks, one of the most successful recurrent neural networks (RNN)-based methods in multivariate time series forecasting and then we introduced DeepLIFT-enabled explanation generation technique. We chose LSTM with the objective that most of the audience can understand our explainable framework. However, our explainability method can be applicable to any deep learning-based forecasting model (i.e., CNN, GRU). Due to the complex architecture and working principle, deep time series forecasting models are very opaque (i.e., black-box) and even AI developers struggle to understand the decision.

We approximate shapley values employing DeepLIFT to track the features' contribution in different layers. We apply DeepLIFT which decomposes the complex LSTM energy forecasting model for a specific prediction by back-propagating to compute the contributions of neurons and approximate the Shapley values to generate explanations. Given that multivariate time series forecasting involves both time and features, we address the challenges of mapping feature contributions with corresponding time frames. Finally, we provide explanations by mapping specific time series and feature importance (i.e., highlighting the contribution of different appliances toward the prediction) with different easily

understandable visualizations. Compared to the conventional application of XAI tools (i.e., LIME [244]) in classification tasks, our method can generate explanations that highlight feature contributions along with their corresponding time frames.

We design a metric to measure the efficiency of the explanations by comparing the highlighted contributions for different appliances toward prediction with the original contributions to the overall consumption. We hypothesize that if the contributions of different features in the explanations for the prediction correlate with the contributions towards the original consumption and have an increasing monotonous relationship, the explanations can be considered effective. The contributions of different appliances toward the overall household consumption can be calculated statistically and represented in vector. Then the contribution vectors for original consumption and prediction are employed to measure effectiveness. If there is a high monotonous correlation between the vectors of contributions for original energy consumption data and Shapley values, we can conclude that the generated contributions using DeepLIFT are analogous. The degree of goodness of the generated explanations can be represented by the correlation coefficient, where the higher the correlation coefficient the better the generated explanations are. However, the major contributions of this research can be summarized as follows:

- We employed DeepLIFT to enhance the explanations for the decision provided by the deep multivariate time series forecasting model by mapping the time and the contribution of different features. Note that our method is applicable to explaining the decisions of any other deep learning-based forecasting models, such as CNN, GRU, etc.
- We designed and introduced an evaluation metric to measure the effectiveness of the generated explanation considering the monotonous relationship between the original and predicted im-

pacts on the overall energy consumption. Our framework achieved high efficiency in terms of the designed metric and can capture appliance contributions toward overall household energy consumption.

- We elicit multiple research gaps in providing human-centered explainability by analyzing existing literature on energy demand forecasting and their explainability. These elicited research gaps would provide future directions for HCI and AI practitioners toward making forecasting systems understandable for users.
- Moreover, the results of multiple experiments on the benchmark datasets with five different households demonstrate that our proposed explainable energy demand forecasting framework achieved effective prediction performance in terms of multiple evaluation metrics.

The rest of the paper is organized as follows: In section 6.2, We summarize state-of-the-art methods in energy demand forecasting and advancements in XAI for time series forecasting. We also highlight the research gaps towards achieving explainable energy demand forecasting. We then introduce our proposed explainable deep household energy demand forecasting framework, called *ForecastExplainer*, which approximates SHAP values using DeepLIFT (section 6.3). The experimental results with multiple settings on two different datasets are analyzed and discussed in section 6.4. We also present the generated explanations and their effectiveness in this section. Conclusions and key findings are presented in section 6.5. Finally, section 6.6 outlines future research directions, focusing on human-centered evaluation of the explanations and eliciting further requirements through an empirical study from a user-centered perspective in the context of smart home systems.

6.2 Literature Review

This section presents an extensive discussion of the prior works on energy demand forecasting and the progress of explainable artificial intelligence, especially for time series forecasting models. Therefore, we present prior research works reviewing published literature in two different sub-sections. At the end of this section, we highlight the research gaps towards making the energy demand forecasting system explainable.

6.2.1 Energy demand forecasting

The methods to forecast households' energy demand can range from classical to complex ML and deep neural networks (DNN)-based techniques [197, 311, 341]. In the recent past, there is a huge interest in applying ML and deep learning techniques to forecast household energy demand [341, 184, 163, 162, 161, 311]. Kazemzadeh et al. [153] proposed a hybrid long-term demand forecasting model based on data mining techniques. They applied particle swarm optimization in the hybrid model consisting of support vector regressor, Auto-Regressive Integrated Moving Average (ARIMA), and Artificial Neural Network (ANN). Similar to Kazemzadeh et al. [153], Yan et al. [330] proposed an LSTM-based hybrid model for modeling individual household energy consumption. They leveraged the stationary wavelet transform (SWT) technique to increase the dimension of the data and tackle the volatility and then applied the LSTM-based deep learning model. Fu et al. [105] proposed a data-driven situational awareness framework that monitors energy consumption on the campus. Their framework consists of two different components including energy demand forecasting models and anomaly detection systems to support immediately on the campus. Similar to our research, their energy demand forecasting system is modeled by LSTM-based neural network architecture. However, our goal in this research is to explain

the prediction by energy demand forecasting models, not having a new forecasting model. Since LSTM is most widely used and successful in energy demand forecasting tasks, we have applied our explainable framework aiming to generate meaningful explanations for particular prediction highlighting different features corresponding to the time duration.

Kim & Cho [164] proposed a CNN-LSTM neural network model combining convolutional neural networks (CNN) and LSTM networks to extract better temporal and spatial features that can predict household energy effectively. A hybrid deep learning framework is proposed by Syed et al. [305] employing a fully connected neural network followed by a unidirectional LSTM and bi-directional LSTM (Bi-LSTM) model to overcome the temporal dependencies of the energy consumption. Chadoulos et al. [57] introduced a deep learning model combining recurrent neural networks (RNN) and multi-layer perceptron (MLP) to forecast hourly demand for different households considering consumers' profiles. In addition, the method can capture the past and future impacts of time series and consumer profiles.

Some prior studies [163, 162, 161, 311] modeled the energy demand forecasting task exploiting DNN, CNN, LSTM and auto-encoder, and explained the prediction. Kim et al. [163, 162, 161] conducted multiple studies and proposed multiple methods for forecasting household electric demand. In the study [161], an auto-encoder-based deep learning model is proposed which can predict the energy demand for each 15, 30, 45, and 60 minutes for various household scenarios. Similar to the previous study, methods presented in [162, 163] applied also an auto-encoder-based model consisting of four different components. The first component models the past energy consumption, and then a subencoder models consumption information and processes as latent variables. The third component maps the future demand considering the latent variables. Lastly, the final component tried to interpret the important elec-

tric information to highlight the model's global interpretability.

Chakraborty et al. [58] introduced explainable artificial intelligence to predict climate change impact on a scenario-based building cooling energy forecasting. For optimal management of the building's energy, Eseye and Lehtonen [93] introduced short-term forecasting of heat energy demand with integrated ML models. Their model incorporated a support vector machine (SVM) with an imperialistic competitive algorithm embedding feature selection technique combining binary genetic algorithm and Gaussian process regression. Zhang et al. [341] proposed an explainable energy forecasting model exploiting AI-based techniques. They trained a surrogate model to mimic the original trained model and interpret the model. Ahmed et al. [8] proposed a random neural network-based energy prediction model for large buildings. A wide range of experiments was conducted on one-year energy data, and they have achieved better performance than artificial neural networks and support vector machine-based regression techniques.

6.2.2 Explainable AI in time series forecasting

Though there is some attention to making the model interpretable in time series forecasting, most of those methods attempted to explain only the algorithmic decision-making to increase the model's performance and debugging [215, 162]. However, the methods for generating explanations for general users are not so common [87, 145]. Assaf & Schumann [27] proposed a gradient-based technique to explain the prediction from a CNN-based time series model. The explanations are provided via a saliency map considering the time dimension and the features. Their method can identify the specific time duration and highlight the most important factors on the time for the particular prediction.

Similarly, Amal et al. [256, 257] proposed a CNN-based explainable time

series forecasting model using adaptive saliency maps explanations. Prior studies [249, 89, 32, 264, 260, 255] surveyed explainable methods on time series data by highlighting the overview, impacts and available methods in the field of explainable models for time series data. Ilic et al. [133] introduced an explainable boosted regression technique for time series forecasting. Their method provides explanations through regression trees. A heatmap-based explainable technique by Kim & Cho [162] is presented to explain the auto-encoder-based forecasting model. Zdravkovic et al. [337] applied local interpretable model-agnostic explanations (LIME) [244] to explain the heat energy demand forecasting model. LIME and Shapley additive explanation (SHAP) [195] based explainable models are also employed for explaining time series classification not forecasting tasks.

The explanations and their representation interface will surely be different in the case of human-centered explanations for general users [216, 247, 239, 145]. Some explainable methods are also published recently where they focused on human-activity recognition and e-health in smart home environments [155, 39, 76, 22, 75].

In conclusion, there is still a big gap in human-centered XAI systems for general smart home users, particularly in the energy demand forecasting problem. In this research, we try to explain the complex energy demand forecasting prediction with approximating shapley values incorporating DeepLIFT. Our visualizations towards explaining specific decisions might help general users so that they can build more awareness of consuming energy in their homes. Moreover, these explanations might help towards optimizing their energy consumption considering the factors behind the predictions.

Table 3: The summary of the state-of-the-art research on household energy demand forecasting with possible research gaps.

Authors and Paper (ref.)	**Summary of Method & Contribution**	**Gaps related to explainability**
Kazemzadeh et al. [153] & Yan et al. [330]	Both papers proposed hybrid models to predict household energy consumption. Kazemzadeh et al. [153] applied ARIMA and ANN, whereas Yan et al. [330] applied a dimensionality reduction approach and LSTM-based DL forecasting technique.	Did not consider explainability
Fu et al. [105]	Applied a data-driven situational awareness for monitoring energy consumption in a university campus	Did not consider explainability
Kim & Cho [164]	Proposed a predictive method combining CNN and LSTM (CNN-LSTM) for extracting better spatial and temporal feature for residential energy demand prediction.	Did not consider explainability
Kim et al [163, 162, 161]	Modeled the energy demand forecasting problem using different neural network-based approaches including CNN, LSTM and auto-encoder.	The proposed models have global interpretability. But the method can not explain for a particular prediction.

Chakraborty et al. [58]	Introduced an explainable AI-driven approach to predict climate change impact for building's cooling energy forecasting.	Incorporated Shapley additive explanations for highlighting the feature impacts. But the method can not explain for a particular prediction.
Eseye and Lehto [93]	Proposed a method for forecasting energy demand for household with several ML models.	Did not consider explainability
Zhang et al. [341]	Introduced interpretable energy forecasting model by developing a surrogate model that might mimic the original model's performance.	Only provide interpretability about the local mechanism of the model
Assaf & Schumman [27]	Proposed a CNN-based explainable time series forecasting model via saliency map/heat map.	The model can provide global interpretability with a heatmap highlighting both time and features. But the method can not explain for particular prediction.
Amal et al. [256, 257]	Proposed adaptive saliency map-based explanation techniques for time series forecasting models including CNN and ensemble classifiers.	The model can provide global interpretability using a saliency map. But the method can not explain for particular prediction.

Ilic et al.[133]	Introduced an explainable boosted regression technique and the explanations can be presented via regression tree.	Explainable only for boosted regression technique.

The summary of notable state-of-the-art methods in energy demand forecasting and explaining time series forecasting is depicted in 3. We observed that most of the studies in energy demand forecasting are not explainable. The studies focusing on explainability in energy demand forecasting only tried to highlight different features for global interpretability. We also include related works [27, 256, 257, 133] that aimed at explaining time series forecasting models. We observed that few methods tried to explain the forecasting model based on time and features. However, they only focus on global interpretability so that AI practitioners can improve the model's performance. Nevertheless, the explanations are highly technical and not easily understandable by the general users in smart homes. To the best of our knowledge, there is no such model that can provide local explanations for energy demand forecasting model highlighting time and features. In this paper, we try to fill the gap in generating understandable explanations for certain predictions highlighting the time and features in an easily understandable way. The primary goal is to provide such explanations to the user so that they can be more optimal and aware when they utilize a particular appliance.

6.3 Explainable Energy Demand Forecasting Framework

This section presents our explainable energy demand forecasting framework. In particular, we have two major components in this framework, (i) a deep LSTM networks-based energy demand forecasting model and (ii) an inference and explanation generation technique by approximating

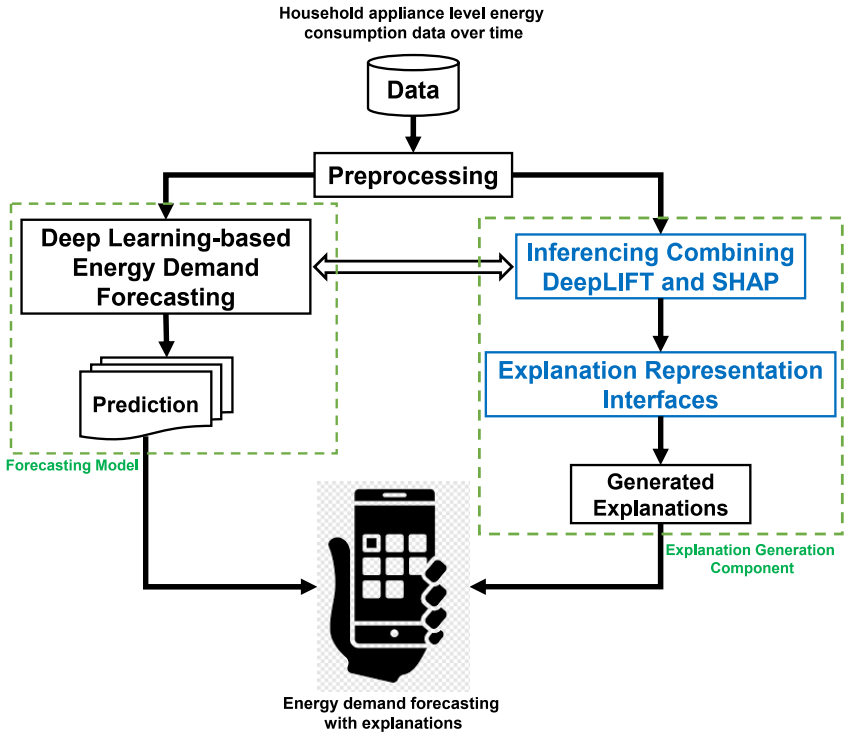

Figure 13: An overview of explainable energy demand forecasting framework for smart home

Shapley values applying DeepLIFT. The high-level building blocks of the explainable energy forecasting framework are illustrated in Fig. 13. In summary, we first preprocess the time series data by handling missing values and filtering out the noisy data. The data were collected with 1-minute granularity. We re-sampled the data applying the sum on an hourly, daily and weekly basis. Then, we trained an efficient deep LSTM network-based energy demand forecasting model that can predict hourly, daily and weekly energy demand in the household. Finally, we apply DeepLIFT to explain the individual prediction approximating Shapley values and provide comparatively understandable explanations through visualization. However, these two components - the forecasting model and explanation generator - are separate from each other and the prediction performance is not affected both positively and negatively. However, the DeepLIFT-based explanation generation techniques map

the impact of features in different layers and can highlight the contribution corresponding to a particular time duration.

With enormous success in predictive modeling, the complex deep neural network (DNN)-based approaches have received huge attention in every sector. Recurrent neural network (RNN) is a successful DNN technique that can model sequential data better. However, traditional RNN faced a problem in memorizing long-term dependency. This is widely called a gradient vanishing problem. To make the presentation of the DeepLIFT-enabled explainable forecasting model simpler for the audience, we chose a long-short term memory (LSTM) network, a widely used variant of RNN that can overcome this long-term dependency problem. For sequential predictive modeling, LSTM is one of the most successful DNN models, especially for multivariate time-series forecasting. However, LSTM has a very complex architecture and hence the decision-making procedure of these types of predictive model are very opaque, even AI practitioners often fail to understand why a particular decision is being predicted. We applied an efficient explainable LSTM networks-based forecasting model for predicting the household energy demand. Therefore, we first discuss the details of our energy demand forecasting framework applying LSTM networks and then we present our explanation generation technique for specific prediction. Note that our DeepLIFT-enabled explanation techniques can be applicable to other RNN-variants (i.e., GRU)-based DL forecasting models.

6.3.1 Energy demand forecasting framework with LSTM networks

In contrast to traditional feed-forward DNNs, LSTM networks possess feedback connections that facilitate the processing of sequential data and the retention of crucial information within the sequence. This capability empowers them to effectively handle subsequent data points. Drawing inspiration from the accomplishments of LSTM-based mod-

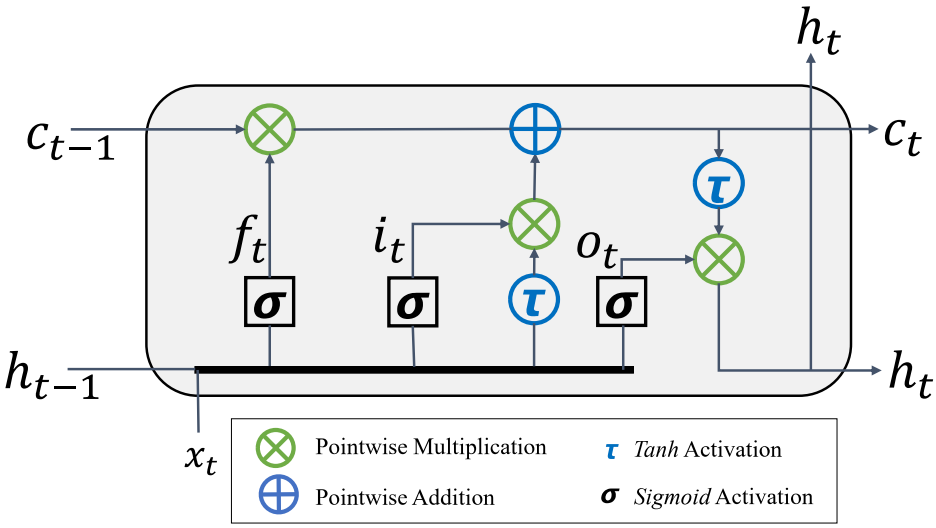

Figure 14: A LSTM block with forget, input and output gates, f_t, i_t and o_t, respectively

els in addressing text, audio, and time series forecasting challenges, we have developed a sophisticated energy forecasting model employing deep LSTM architecture, featuring multiple LSTM layers. Furthermore, the technique employed to generate explanations for such a successful model holds potential for broader application across diverse domains. It is, however, a reasonable expectation that heightened model performance correlates with increased network depth. Guided by this rationale, we meticulously fine-tuned our deep LSTM model, making meticulous selections of optimal parameters, including the number of hidden layers, the number of hidden units within each LSTM layer, and the size of epochs. In our tiered network, the output of the $(k-1)$-th LSTM layer is harnessed as the input for the subsequent k-th layer. This intricately layered architecture empowers our model with the capacity to make predictions regarding future energy demand.

To address the issue of gradient vanishing, LSTM cells incorporate three

distinct components, known as gates, at a specific time step. The three gates include the forget gate, input gate, and output gate. These gates serve the purpose of regulating the information flow that enters, remains stored, and exits the network, respectively. Each of these gates has its own neural network, functioning as a filter within the LSTM cell. It is important to note that the output of an LSTM cell is dependent on the current input data, the current long-term memory, and the previous hidden state. The diagrammatic representation of an LSTM block, showcasing the functioning of its different gates and states, is depicted in Fig. 14. Let the current input data be x_i at time t and the previous hidden state be h_{t-1}. For each gate, we denote input weight as U, recurrent weight as W and bias as b. First, LSTM processes the *forget gate* at time t, f_t which is the neural network working as follows:

$$f_t = \sigma(X_t \mathbf{U^f} + h_{t-1} \mathbf{W^f} + b_f), \tag{6.1}$$

where forget get apply a *sigmoid* activation function that returns output in the interval [0,1] and σ represents the *Sigmoid* activation function. When the output of this network is close to 1, the forget gate chooses the input component as relevant. Otherwise, it neglects for output closer to 0 as irrelevant.

After throwing out the irrelevant information, LSTM next decides which information is to store and update the cell state. For this, it employs an input gate as follows:

$$i_t = \sigma(X_t \mathbf{U^i} + h_{t-1} \mathbf{W^i} + b_i). \tag{6.2}$$

A *tanh* activation function-based layer is then applied to combine the previous hidden state and new input data for generating a new memory update vector $\tilde{C}_t$ as follows

$$\tilde{C_t} = tanh(X_t \mathbf{U^c} + h_{t-1} \mathbf{W^c} + b_c). \tag{6.3}$$

Applying point-wise multiplication and addition, the old state C_{t-1} is then updated as C_t.

$$C_t = C_{t-1} \otimes f_t \oplus i_t \otimes \tilde{C_t}. \tag{6.4}$$

With the help of output get, LSTM finally produces the output as the next hidden state by processing cell state and input

$$o_t = \sigma(X_t \mathbf{U^o} + h_{t-1} \mathbf{W^o} + b_o). \tag{6.5}$$

To produce the final state, a point-wise multiplication is applied between o_t and cell state passed through a *tanh* activation function.

$$h_t = o_t \otimes tanh(C_t). \tag{6.6}$$

6.3.2 Explaining predictions with DeepLIFT approximating the Shapley Value

To explain the opaque and very complex LSTM-based energy demand forecasting model, we approximate Shapley values employing DeepLIFT (Deep Learning Important FeaTures). DeepLIFT is a method that decomposes the complex deep neural network-based methods for specific prediction by back-propagating to compute the contributions of neurons. For a given prediction, this method provides local explanations summarizing the contributions computing the "*difference in output from some reference output considering the difference in input from some reference input*" [293].

Let us assume that we have a neural network prediction model with an input layer with neurons $\{x_1, x_2, x_3, \cdots x_n\}$, some hidden layers with sets of neurons $\{h_1, h_2, h_3, \cdots, h_n\}$ and a target output neuron t. Consider $f(x)$

to be the activation of a particular neuron and $f(x')$ to be the reference activation. DeepLIFT calculates the contributions scores as follows:

$$\Delta t = f(x) - f(x') = \sum_{i}^{n} C_{\Delta x_i \Delta t}, \tag{6.7}$$

where $C_{\Delta x_i \Delta t}$ refers to the contribution score in each neuron x_i. For a given target output t and its reference activation t^0, the difference-from-reference is computed as $\Delta t = t - t^0$. Eq. 6.7 is also called a *summation-to-delta* property.

Let Δx be the difference-from-reference of any input neuron calculated in the same procedure described previously. For target output t and difference from output reference Δt, we can define the multiplayer by averaging the difference as follows:

$$m_{\Delta x \Delta t} = \frac{C_{\Delta x \Delta t}}{\Delta x}. \tag{6.8}$$

It can be seen as a contribution of Δx to the target difference Δt, computed by diving with Δx.

Let $\{x_1, x_2, x_3, \cdots, x_n\}$ be the set of neurons for a complete neural network, $\{h_1, h_2, h_3, \cdots, h_n\}$ be the hidden layers with neurons' set and t is a target output neuron, we can define the contribution multiplier as a chain rule:

$$m\Delta x_t \Delta t = \sum_{j} m_{\Delta x_i \Delta h_j} \cdot m_{\Delta h_j \Delta t}, \tag{6.9}$$

where the contribution is calculated by applying iterative chain rules in each layer. This can be applicable to any number of layers in the networks. By applying this chain rule, the contribution in terms of multipliers can be computed for a given target employing back-propagation. This is analogous to the chain rule in partial derivatives.

We employ DeepLIFT to approximate the Shapley values to explain any particular prediction in energy demand forecasting. Here, the multipli-

ers are represented in terms of SHAP values ϕ_i:

$$m_{x_j,f_j} = \frac{\phi_i(f_j,x)}{x_j,\mathbb{E}[x_j]} \tag{6.10}$$

Similar to the chain rule mentioned in Eq. 6.9, this can be defined as follows:

$$m_{x_j,f_j} = \sum_j m_{x_j,h_j} \cdot m_{h_j,f_j} \tag{6.11}$$

Here, we approximate the reference value by averaging over the background instances. The approximation is done by summing up the difference between the expected model output of the background model and the output of the current model, $f(x) - \mathbb{E}[x_j]$. The SHAP values are computed as:

$$\begin{aligned}\phi_i(f,a') = \sum_{z' \subseteq \{a'_1,a'_2,\ldots,a'_n\} \setminus \{a'_i\}} & \frac{(|z'|)!(M-|z'|-1)!}{M!} \\ & \cdot [f(z' \cup a'_i) - f(z')],\end{aligned} \tag{6.12}$$

where a is the features vector and z' and a subset of the features employed by the model f. $f(z')$ is the prediction by the model f.

a' is the vector with feature values to be explained and can be defined as $[f(z' \cup x) - f(z')]$ and M is the number of features. The prediction by the model f is denoted by $f(z')$. Moreover, SHAP values are computed by a standard game-theoretical approach and utilised Shapley values to have a unified interpretable model with fast computation. More mathematical and technical details of DeepLIFT and SHAP can be found in the study published by [195] and [293], respectively.

6.4 Experiments and Evaluation

This section presents the details about datasets, evaluation metrics, experimental settings, performance in energy demand forecasting and the

Table 4: Description of different variables in EnergyData

Feature, f_i	**Description**
Global active power, f_1	Household global minute-averaged active power
Global reactive power, f_2	Household global minute-averaged reactive power
Voltage, f_3	Minute-averaged voltage (in ampere)
Global Intensity, f_4	Household global minute-averaged current intensity
Sub-metering_1, f_5	It corresponds to the kitchen, containing mainly a dishwasher, an oven and a microwave (hot plates are not electric but gas-powered)
Sub-metering_2, f_6	It corresponds to the laundry room, containing a washing machine, a tumble drier, a refrigerator and a light
Sub-metering_3, f_7	It corresponds to an electric water heater and an air-conditioner

generated explanations.

6.4.1 Datasets

We conducted experiments on two public benchmark datasets on household electric energy consumption including EnergyData,[5] and REFIT data [217]. Here we present the summary of two different datasets.

Household energy consumption dataset (EnergyData): The data has been from a house for 47 months, particularly from December 2006 until November 2010. The dataset consists of different energy consumption measures including global active power, global reactive power, global intensity and consumption in different household areas. A brief description of each feature is summarized in Table 4. *Submetering_1* represents the active power consumed by multiple appliances including a dishwasher, an oven, and a microwave. The active power consumption by the laundry room containing appliances including a washing machine,

[5] `https://archive.ics.uci.edu/ml/datasets/individual+household+electric+power+consumption`

a tumble-drier, a refrigerator, and a light is represented by *submetering_2*. The power consumed combined by an electric water heater and an air-conditioner is denoted as *submetering_3*. All the above-mentioned measures were collected for every minute.

REFIT smart home dataset:

This dataset contains cleaned electrical consumption data for 20 households with different properties [217]. The dataset includes the aggregate electricity consumption and appliance-level consumption for 9-10 home appliances in watts at an 8-second granularity. It was collected as part of the REFIT project[6]. For our experiments, we selected four diverse households based on different property characteristics as listed in 5. The selected households are House 2, House 5, House 8, and House 13. The properties of each household, including the number of occupants, construction year, total owned appliances, type of the house, and size in terms of bedrooms, are summarized in 5. Additionally, we provide a list of appliances from which the energy consumption data were collected for each respective house in 6.

Table 5: Properties of the selected households

House	**House Information**				
	Occupants	**Year**	**Appliance**	**Type**	**Size**
House 2	4	-	15	Semi-Detached	3 bed
House 5	4	1878	44	Mid-terrace	4 bed
House 8	2	1966	35	Detached	2 bed
House 13	4	2002	28	Detached	4 bed

6.4.2 Evaluation metrics

We employed different evaluation metrics to validate the performance of our method in forecasting household energy demand. Out of numerous evaluation metrics, we employ four metrics including mean absolute er-

[6]Personalized Retrofit Decision Support Tools for UK Homes using Smart Home Technology', Grant Reference EP/K002368/1/1

Table 6: The list of appliances considered in collecting data for the selected households

House #	**List of appliances**
House 2	Fridge-Freezer, Washing Machine, Dishwasher, Television, Microwave, Toaster, Hi-Fi, Kettle, Overhead Fan
House 5	Fridge-Freezer, Tumble Dryer, Washing Machine, Dishwasher, Desktop Computer, Television, Microwave, Kettle, Toaster
House 8	Fridge, Freezer, Washer Dryer, Washing Machine, Toaster, Computer, Television, Microwave, Kettle
House 13	Television Site Freezer, Washing Machine, Dishwasher, Network Site, Microwave, Microwave, Kettle

ror (MAE), mean absolute percentage error (MAPE), mean squared Error (MSE), and root mean squared error (RMSE) for evaluating the performance. To measure the effectiveness of the generated explanations by DeepLIFT, we introduced a new metric named *contribution monotonicity coefficient, CMC*.

Evaluation metrics for forecasting:

MAE. The average of absolute differences between predicted values and the original values are referred to as mean absolute error (MAE), which can be computed as follows:

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}, \tag{6.13}$$

where y_i denotes the foretasted values and x_i is the original energy consumption. This metric calculates the degree of average error made by the predictive model. The low MAE values close to zero indicate the high accuracy of the predictor. Since this is the arithmetic average, it can be affected by sampling fluctuation.

MAPE. This can be computed by dividing the absolute difference between predicted and original values by original value.

$$MAPE = \frac{100}{n} \sum_{t=1}^{n} \left| \frac{x_t - y_t}{x_t} \right| \tag{6.14}$$

To address the sampling fluctuation issue, the division is done by x_t for corresponding predicted and original values.

MSE. This is another widely used evaluation metric that calculates the average error by applying the squared difference between the predicted and original values instead of the absolute difference.

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (x_i - y_i)^2 \tag{6.15}$$

One of the features of this metric is that it penalizes outliers and/or large errors more than minor differences because of employing a square function. Compared with MAE and MAPE, this evaluation metric is better as it overcomes the extreme and zero value problem.

RMSE. As an extension of MSE, RMSE applies the square-root function over the squared difference between original and predicted values.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{y}_i)^2} \tag{6.16}$$

This metric makes it easier to understand the performance of any forecasting model than other metrics. However, for all metrics, the lower the value of any metric, the better the performance of the forecasting model is.

Metric to measure explainability: We present the explanations generated by approximating Shapley values with DeepLIFT using different visualizations to comprehend specific predictions from our forecasting

methods. The core of these explanations is a list of contributions from various appliances and features toward the predicted future overall consumption. To assess the quality of these explanations, we compare the highlighted contributions for different appliances in households with the original contributions in the test data. We statistically compute the contributions of different appliances to total energy consumption. In other words, how much a particular appliance is responsible for the overall energy consumption for a certain time? The explanations we generate also depict contributions from different appliances towards future consumption, expressed as approximate Shapley values computed by applying DeepLIFT to the deep forecasting model.

Our hypothesis for evaluating the computed contributions is to examine how monotonous and correlated the generated contributions are with the original contributions to overall energy consumption. If there are high monotonous correlations between the vectors of contributions for original energy consumption data and Shapley values, we can conclude that the generated contributions using DeepLIFT are analogous. The degree of goodness of the generated explanations can be represented by the correlation coefficient, where a higher correlation coefficient indicates better-generated explanations.

Following the above-mentioned hypothesis and intuition, we first compute the total energy consumed by a particular appliance A_i, denoted as T_{A_i}, and then calculate the contribution of the appliance C_{A_i} by dividing the total energy consumption of the household T_H by T_{A_i} ($C_{A_i} = \frac{T_{A_i}}{T_H}$). Using the same formula, we compute the contributions for all appliances represented in a vector. On the other hand, we have a contribution vector S_{A_i} in terms of Shapley values by DeepLIFT, representing the predicted contributions for different appliances toward overall predicted consumption. For two given contribution vectors, we can compute the correlation coefficient between them. To do this, we consider the Spear-

man correlation coefficient to assess the correlation between original and predicted contributions to overall energy consumption. The Spearman Rank-correlation coefficient is chosen because it can identify the monotonous relationship between two vectors. If we find a high correlation and an increasing monotonous relationship, we can conclude that the predicted contributions are analogous to the original contributions.

Contribution monotonicity coefficient, CMC: Given two contribution vectors, $C = \{C_{A_1}, C_{A_2}, C_{A_3}, \cdots, C_{A_n}\}$ and $S = \{S_{A_1}, S_{A_2}, S_{A_3}, \cdots, C_{A_n}\}$ that represent the normalized contributions for different features. The C_{A_i} denotes the real contribution towards the overall consumption and the S_{A_i} represents the predicted contributions by DeepLIFT techniques in terms of Shapley values. We compute the Spearman-ranked correlation coefficient-based measure *contribution monotonicity coefficient*, CMC as follows:

$$\rho_{CMC(C_A,S_A)} = 1 - \frac{6\sum d_i^2}{n(n^2-1)}, \tag{6.17}$$

where d_i is the difference between the ranks of the contribution score C_{A_i} and S_{A_i} and n is the length of the vectors. The higher the value of ρ, the better the explanation is. The positive ρ indicates the increased monotonous relationship between the vectors and the negative indicates decreasing.

6.4.3 Experimental setting

The data were collected by measuring the energy consumption in different household areas and appliances for 1 minute and 8 seconds time intervals for EnergyData and REFIT datasets, respectively. We converted the consumption in the datasets in three different forms applying resampling in an hourly, daily and weekly manner. We have summed up the energy consumed hourly, daily and weekly. Then we applied a classi-

cal MinMax scaller to transform every feature's values in similar ranging from zero (0) to one (1). Along with this, we also have a de-scaler function so that we can convert the predicted values (energy forecasting) to original units.

However, we applied 85% of the samples as a training set and the 15% samples left were used for testing the model in all three types of forecasting namely, hourly, daily and weekly. For hourly forecasting, in total, we have household energy consumption data for 34589 hours (i.e., for EnergyData). The sequence lengths for three different forecasting models were 24, 30 and 7, respectively. For all models the number of input features in each sample was the same. Along with regular features including the consumption by different household areas and appliances, we also employ days of the week, month of the year, and quarter of the year as features. Other than that, we subtracted the summation of energy consumed by three different areas (sub-meters) from the total household energy consumption for EnergyData and used it as a new feature. We conducted experiments for both datasets by applying 5-fold cross-validation and applied arithmetic averages to calculate the performance in terms of different evaluation metrics. The details of the parameters of our LSTM models are summarized in Table 7.

Table 7: Summary of the parameters of the LSTM-based forecasting model

Parameter	**Hourly**	**Daily**	**Weekly**
# of features	10	10	10
Sequence Length	24	30	7
# of hidden layers	2	2	2
# of hidden units	64	64	64
# of Epoch	100	100	50
Learning rate	0.001	0.001	0.001
Batch Size	1024	1024	64

After training our LSTM-based forecasting model with adequate training data, we applied our inference component to identify the facts (i.e., the

importance of different features corresponding to time). Our inference and explanations interface then visualize the impact of different features with time duration in different forms. This explanation generation component is different from the demand forecasting model. DeepLIFT can identify contributions by approximating shapely values for different features by mapping the changes of the values in the different layers. Finally, we highlight different features' contributions with time duration. To compare the performance of our explainable forecasting framework, we applied the method proposed by Kim et al [164]. They applied a CNN-LSTM-based deep learning model. We designed and conducted the experiments by applying their method with the same feature scaling and normalization techniques.

6.4.4 Experimental results

Along with the features described in Table 4, we also extracted handcrafted features and introduced four different features including seasonality. Generally, the daily energy consumption is dependent on the type of days. It is expected that the overall energy consumption on the weekend is supposed to be different than on the weekdays. Similarly, the season has a great impact on the overall consumption, i.e., the daily consumption in the winter season will be different from the consumption in summer and the consumption trend will be different in autumn as well. Therefore, we extract three new features namely, *the day of the week*, *the month of the year*, and *the quarter of the year*. Other than these features, we noticed that the total energy measures by three submeters are smaller than the total energy consumption. Therefore, we add another new feature named *others* that indicates the energy consumption extracted by subtracting the summation of three sub-meters from the total energy consumption. We conducted experiments to predict hourly, daily and weekly energy demand to validate the performance

of our framework.

Table 8: The performance of our proposed explainable forecasting compared to other methods.

Mode	Method	MAE	MAPE	MSE	RMSE
Hourly	**Our Framework**	**0.075**	**48.9**	**0.009**	**0.096**
	Kim et al. [164]	0.077	77.5	0.009	0.098
	Linear Regression	0.5022	83.74	0.4247	0.6517
Daily	**Our Framework**	**0.052**	**23.3**	0**.005**	**0.069**
	Kim et al. [164]	0.063	28.4	0.006	0.083
	Linear Regression	0.3915	52.69	0.2526	0.5026
Weekly	**Our Framework**	**0.119**	27.7	**0.019**	**0.138**
	Kim et al. [164]	0.121	**26.4**	0.021	0.146
	Linear Regression	0.3199	41.33	0.1480	0.3847

As we noted earlier, the main objective of our method is to explain the complex forecasting model applying DeepLIFT to approximate the Shapley values that highlight the contribution of different features corresponding to time. Nevertheless, the performance of our framework in forecasting hourly, daily, and weekly energy demand is summarized in Table 8. We conducted experiments by applying 5-fold cross-validation and applied arithmetic average to calculate the metrics. Along with our framework, we also reported the performance of other well-known household energy forecasting models. We can see that the performance of our method is quite consistent and outperformed in predicting the energy demand in the household energy demand.

Table 8 highlighted the performance comparison of our demand forecasting frameworks with some known related works including linear regression and a demand forecasting model by Kim et al. [164] that applied a CNN-LSTM-based deep learning model. We conducted experiments following the proposed model applying the same normalization and scaling techniques. The results show better performance than their approach except in terms of MAPE for weekly prediction. Though the performance difference is not significant (27.7 vs 26.4). However, we can see from

the table that the comparison illustrated a consistent performance in forecasting in terms of multiple evaluation metrics.

Along with predicting the total energy consumption, we also carried out experiments to see how our framework performed in predicting consumption in a specific area of the household. Since we have the dataset for three different sub-meters where the energy consumption was measured in the kitchen (i.e., dishwasher, an oven, and a microwave), the laundry room (i.e., containing washing machine, a tumble-drier, a refrigerator and a light), and another room containing water heater and air-conditioner. The performance in predicting energy consumption for specific areas on an hourly, daily, and weekly basis is presented in Table 9. In turn, our framework achieved efficient performance since the prediction errors in terms of each evaluation metric are minimal.

Table 9: Prediction performance in forecasting energy demand for different household areas

Forecasting	**Mode**	**MAE**	**MAPE**	**MSE**	**RMSE**
Submetering_1	Hourly	0.149	0.934	0.028	0.168
	Daily	0.149	0.646	0.031	0.175
	weekly	0.106	0.244	0.020	0.142
Submetering_2	Hourly	0.150	0.912	0.029	0.170
	Daily	0.164	0.677	0.036	0.189
	Weekly	0.140	0.325	0.031	0.177
Submetering_3	Hourly	0.239	1.479	0.120	0.347
	Daily	0.116	0.495	0.022	0.149
	Weekly	0.098	0.230	0.015	0.121

The summary of the experimental results compared to two different forecasting methods including linear regression and deep learning models demonstrated the efficiency of our explainable energy demand forecasting model using LSTM. Moreover, the effectiveness of LSTM in time series forecasting is widely known as state-of-the-art in multiple application areas, which concludes the consistency. However, the performance difference is higher than the other baselines in all evaluation metrics. In

turn, the prediction performance for the kitchen, an important energy consumption household area for hourly, daily, and weekly consumption is quite consistent and got better performance in all evaluation metrics.

To visualize the prediction performance of our framework more explicitly, we presented the predicted hourly energy consumption of our framework compared to the ground truth, actual energy consumption. We presented the hourly prediction for 300 random hours in Fig. 15. We can see that for the maximum data points, our prediction framework performs with great consistency except for a few sudden fluctuations in actual energy consumption hours.

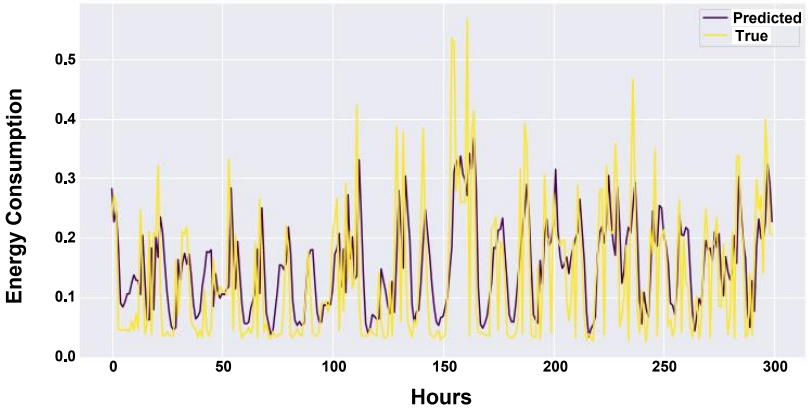

Figure 15: Hourly prediction of our framework compared to the original consumption. The X-axis represents the hours and the Y-axis represents the original hourly energy consumption and predicted energy demand.

6.4.5 Explaining forecasting

The explanations for daily total energy demand forecasting in the household are presented in Fig. 16. We first illustrate the impacts of different household areas to conclude which set of appliances has more responsibility for particular forecasting. Then, we visualize the contributions of seasonality features that reflect the impacts of weather conditions on the final predicted energy consumption. Finally, the contributions

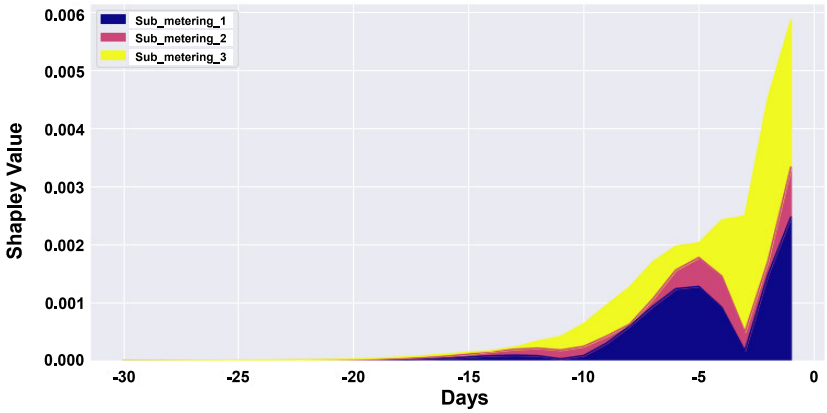

Figure 16: The impact of the consumption in the different household areas on the total energy consumption prediction corresponding to time. The X-axis represents the time and the Y-axis represents the contributions of household areas in terms of Shapley values.

of all mentioned features are presented combinedly in different forms of visualizations. We can see that the daily total energy consumption has a strong impact on energy consumption by air-conditioning and water heaters. The next household area with appliances that have a high impact on the household for overall consumption is the laundry room containing appliances including a washing machine, a tumble-drier and a refrigerator. We can also see the impact of time (in the day) that has impact of consumption in different areas. The impacts in previous days are widely different. On weekends, the impacts were comparatively lower than on regular days..

In turn, we try to see that seasonal impact in the forecasting. Fig. 17 illustrates the explanation in terms of seasonal impact. The figure demonstrates that the *quarter of the year* has the highest impact on the final prediction. It makes sense that the quarter of the year, particularly winter, summer and autumn is supposed to have to higher impact on the energy consumption in households. Similarly, particular months and particular days also have an impact on energy consumption. For example, energy consumption on weekend and weekday are supposed to be

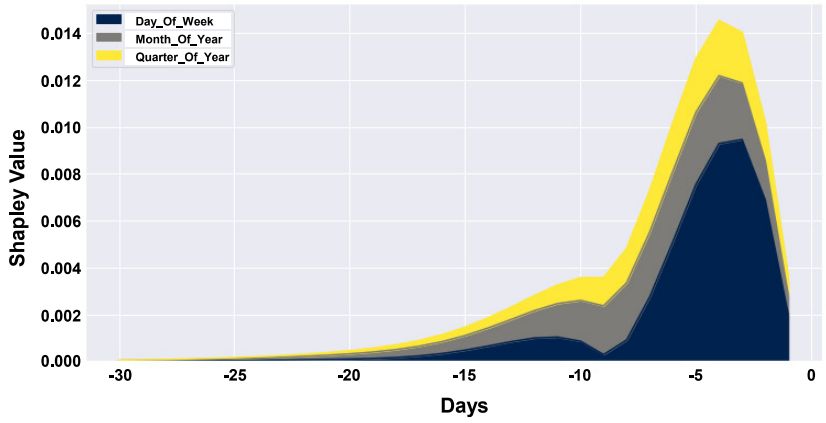

Figure 17: The seasonal impact on the total energy consumption prediction corresponding to time. The X-axis represents the time and the Y-axis represents the contributions of seasonality features in terms of Shapley values.

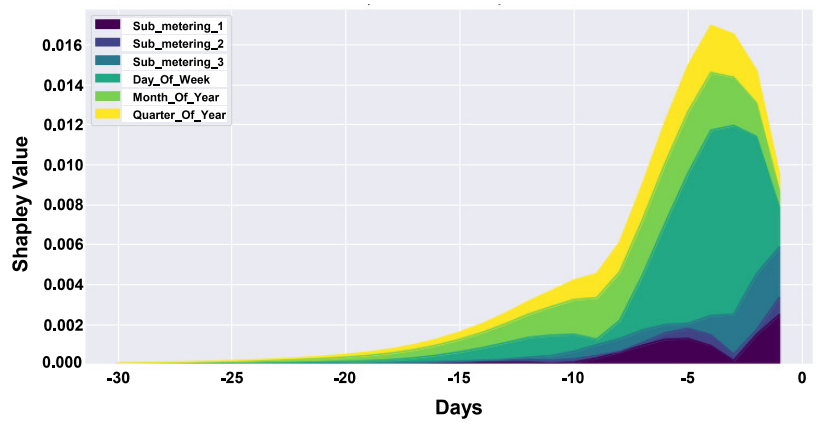

Figure 18: Explanations in terms of the impact of all features corresponding to time. The X-axis represents the time and the Y-axis represents the contributions of features in terms of Shapley values.

different. The month and quarter of the year have a high impact on the energy demand forecasting.

Overall, the explanation for daily prediction is visualized in Fig. 18 in terms of all features corresponding to time. To have better visualization and illustration, the same explanation is presented in different formats in Fig. 19 and 20. Presenting this explanation in an easier way to under-

stand would enable users to become more aware of consuming energy in the household. Moreover, with this explanation, users might think of changing their energy use behavior and patterns to save more household energy, hence leading to a decrease in overall carbon footprint.

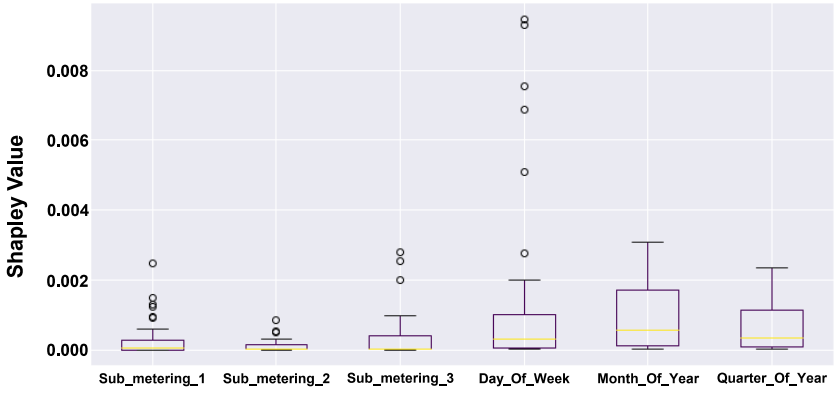

Figure 19: The global importance of different features presented as explanations in terms of Box Plot

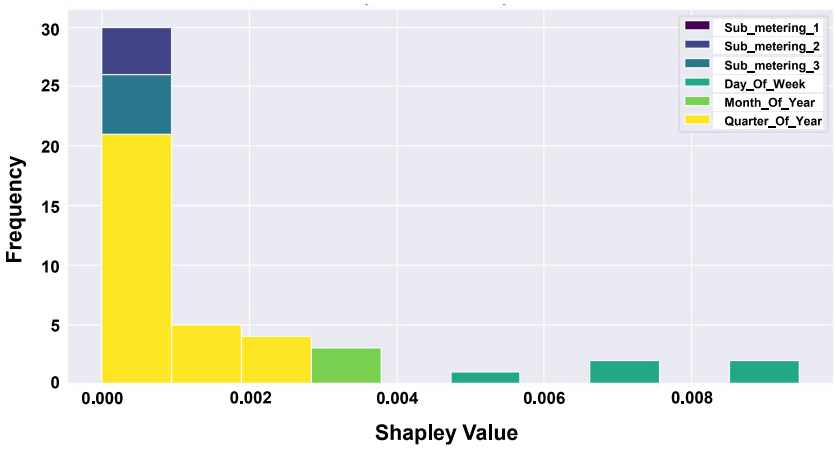

Figure 20: Explanations in terms of histogram highlighting the impact of all features corresponding to time

6.4.6 Performance Robustness

To validate the performance of the forecasting framework, we conducted experiments with another dataset referred to as the “REFIT smart home

dataset". Moreover, we also visualized the explanations to illustrate the appliances having impacts corresponding to the times.

Table 10: Prediction performance in forecasting energy demand on 4 households of REFIT dataset (sec. 6.4.1)

House #	Mode	MAE	MAPE	MSE	RMSE
House 2	Hourly	0.047	2.125	0.007	0.083
	Daily	0.151	0.701	0.037	0.192
	Weekly	0.134	0.557	0.03	0.174
House 5	Hourly	**0.022**	0.516	**0.001**	**0.034**
	Daily	0.085	0.476	**0.012**	0.109
	Weekly	0.076	**0.396**	0.01	0.1
House 8	Hourly	0.037	**0.436**	0.004	0.065
	Daily	0.173	**0.411**	0.041	0.203
	Weekly	0.19	0.464	0.046	0.214
House 13	Hourly	0.044	0.603	0.006	0.079
	Daily	**0.037**	0.698	0.003	**0.056**
	Weekly	**0.041**	0.812	**0.003**	**0.057**

Forecasting performance on REFIT data: The performance of our explainable energy demand forecasting system is illustrated in Table 10. We can observe that the forecasting performance for hourly, daily, and weekly aggregate energy consumption across different households in the REFIT dataset remains consistent across all evaluation metrics. Moreover, compared to the performance on the previous dataset, we can see that the performance, based on all evaluation metrics, is even better. The bold real numbers in the table indicate the best results achieved across all four households for different forecasting modes (hourly, daily, and weekly).

For hourly forecasting, with the exception of MAPE, we can observe that the forecasting performance is better for House 5 across all evaluation metrics. On the other hand, for daily forecasting, our method achieved the best performance for House 13 in terms of MAE and RMSE, House 8 in terms of MAPE, and House 5 in terms of MSE. The weekly forecasting performance is quite similar to the daily performance, showing

better results across all metrics except MAPE for House 13. However, the performance difference across all households is not substantial.

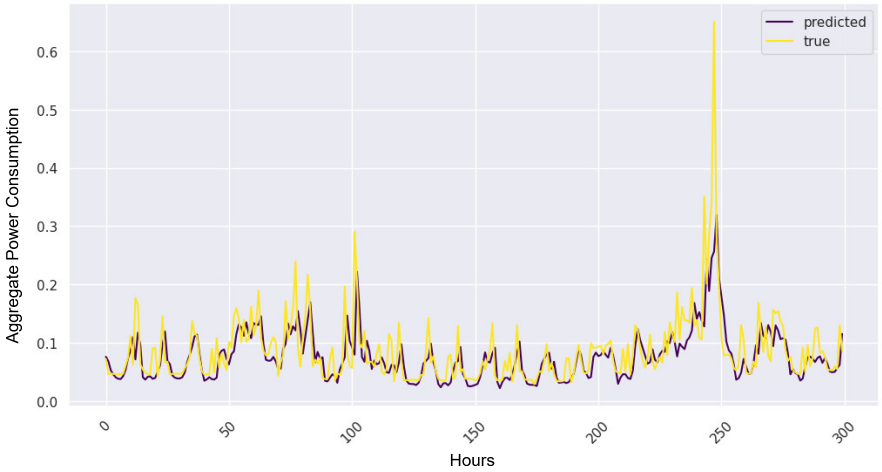

Figure 21: Hourly prediction of our framework compared to the original consumption on house 8 (REFIT dataset). The X-axis represents the hours and Y-axis represents the original hourly energy consumption and predicted energy demand.

Similar to the previous dataset, we present the hourly predictions for house 8 compared to the original consumption in Fig. 21. The X-axis represents 300 random consecutive hours, and the Y-axis represents the normalized aggregate consumption. The figure demonstrates that our method can accurately forecast future consumption over an extended period, except for a very sudden fluctuation near hour 248.

Explaining forecasting on REFIT data: The explanations for house 8 are illustrated in Fig. 22 and 23 in terms of area plot and bar chart. The X-axis represents days and the Y-axis represents the contributions/impacts of different appliances with seasonality. We can see that the most influential appliances are the Fridge, Toster, Kettle, Microwave, etc. In terms of seasonality, *Quarter_of_Year* has the highest influence on the weekly prediction. The contributions or impacts of different feature appliances are illustrated in Fig. 24 in terms of the box plot. We

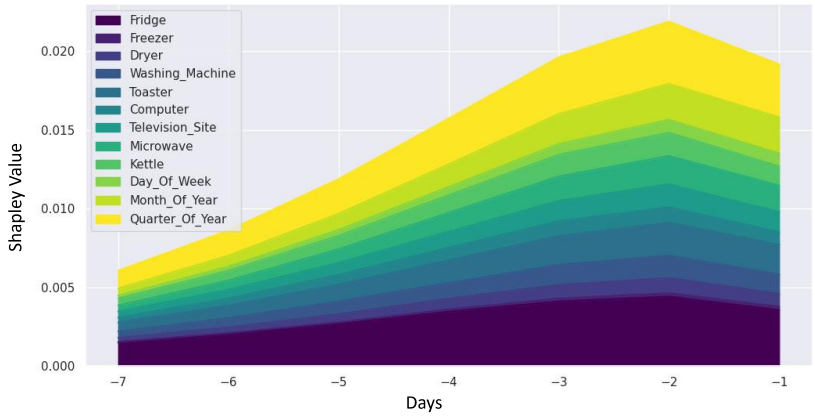

Figure 22: Contributions of different appliances and features corresponding to times (days) in house 8 towards overall weekly aggregate forecasting.

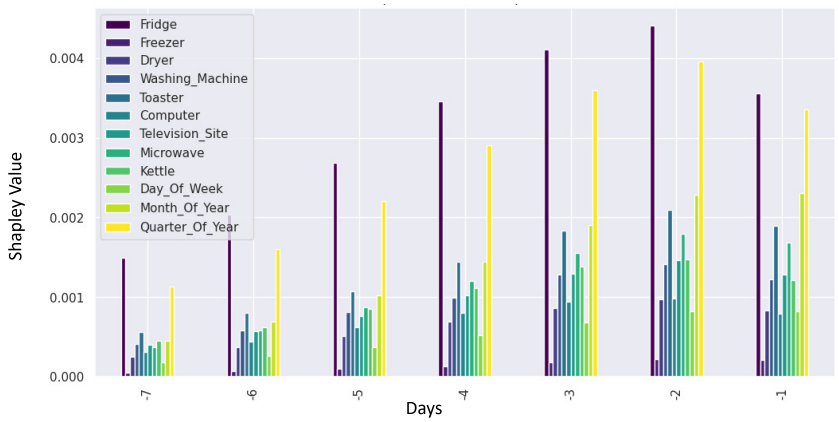

Figure 23: Contributions of different appliances corresponding to times (days) in house 8 towards overall weekly aggregate forecasting.

can see that the Fridge and Toaster are the two appliances having the highest contribution toward the model's prediction.

In general, kitchen appliances such as the Toaster, Microwave, and Kettle collectively have a more significant impact on the overall energy consumption prediction. However, evaluating the generated explanations is subjective, especially in the context of time series forecasting, where explanations become two-dimensional, making quantitative assessment more challenging.

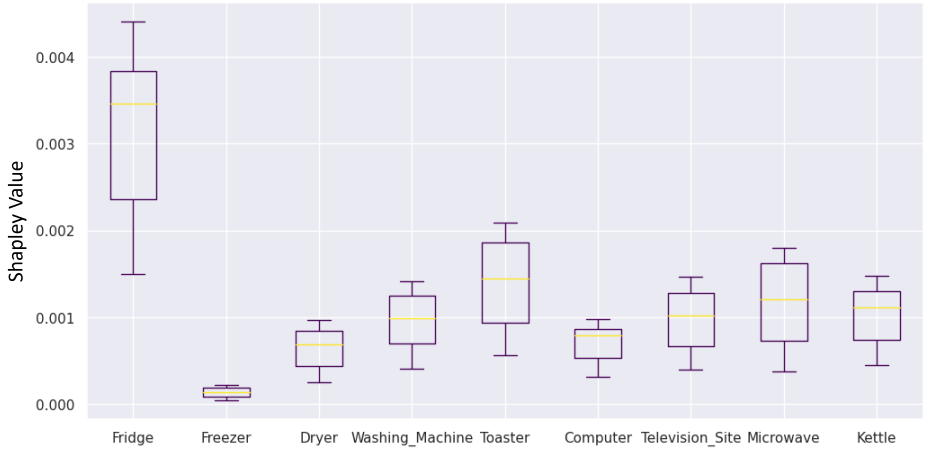

Figure 24: Contributions of different appliances in house 8 towards overall weekly aggregate forecasting.

Table 11: The effectiveness of the generated explanations by DeepLIFT on energy demand forecasting

Dataset	**Household**	**Mode**	ρ_{CMC}	***P*-value**
EnergyData (sec. 6.4.1)	House 1	Daily	0.8857	0.0188
		Weekly	0.7881	0.0318
REFIT Data (sec. 6.4.1)	House 2	Daily	0.8166	0.0072
		Weekly	0.7829	0.0198
	House 5	Daily	0.6985	0.0345
		Weekly	0.7315	0.0280
	House 8	Daily	0.7133	0.0470
		Weekly	0.7315	0.0102
	House 13	Daily	0.7354	0.0297
		Weekly	0.7918	0.0178

6.4.7 Evaluation of generated explanations

We already discussed and reported the forecasting performance on two different datasets which includes data for five different households. We also illustrated the generated explanations in terms of Shapely values approximated with DeepLIFT. We have seen that the explanations can indicate the factors and appliances associated with future consumption. However, in this section, we computationally report the efficiency of our generated explanations for forecasting in terms of *contribution monotonic-*

ity coefficient, CMC. We designed this evaluation metric that can measure the degree of monotonicity coefficient considering the impacts of different appliances on overall energy consumption on original data and predicted consumption.

We presented the performance of explanations for daily and weekly forecasting in terms of *CMC* for both datasets, encompassing all five households, as shown in Table 11. It can be observed that the explanations for daily energy demand forecasting in the EnergyData dataset achieved the highest Contribution Monotonicity Coefficient (CMC) at 0.8857, with an associated lower p-value, numerically 0.0188 ($<$0.05).

Concerning weekly energy demand forecasting, the effectiveness of the explanations for Household 13 in the REFIT dataset attained the highest CMC score. However, for other households, the correlation between the impacts of different appliances on original consumption is highly consistent with the generated explanations. Except for the explanations for daily prediction for House 5, where the monotonicity coefficient exceeded 70%, indicating a consistently increasing relationship between the impacts of appliances computed by DeepLIFT and the original impacts. The performance on three households data including EnergyData, House 2, and 13 even achieved nearly or more than 80% efficiency in terms of CMC with lower p-value. The high correlation between the impact of appliances in explanations and the original overall consumption of the households indicates the efficiency of the generated explanations that might make user sense of any prediction from a deep learning-based forecasting model.

We also compared the impacts of different appliances on future consumption with the findings from the experiments conducted by Stankovic et al.[302], who investigated the contribution of daily activities to overall energy consumption. They used the same dataset as ours and reported that cooking contributes to 16% of the total energy con-

sumption in House 8. In our generated explanation, we also detected the activity of cooking through the consumption patterns of kitchen appliances like the Toaster, Kettle, and Microwave, and their collective contributions, as shown in the box plot (Fig. 24), are correlated. Combining the contributions of these three cooking appliances shows that cooking has the highest impact on overall consumption.

Stankovic et al. [302] further reported that the next significant activities impacting energy consumption are laundering (4%) and watching TV (1%). The laundry activity was detected through the consumption of the washing machine and tumble dryer. Our generated explanation aligns with this too, as we observed that the washing machine and dryer collectively have the second-highest contributions, and the television_site also makes considerable contributions.

Based on the careful analysis on the evaluation of the explanations computationally, we can conclude that our method's generated explanations effectively identify the impact of different appliances on energy consumption. With further empirical studies involving smart home users, we aim to enhance and validate the explanation quality, enabling users to optimize their energy consumption effectively. The predicted future energy demand and the explanation can help the users with energy consumption literacy [268, 270] and the policymaker can think of adopting our method in dynamic pricing and energy policy optimization.

6.5 Conclusion

This paper presents an explainable energy demand forecasting system where we attempt to generate easy-to-understand explanations for forecasting decisions for smart home users. For doing so, we approximate the SHAP values by applying DeepLIFT to identify the feature's contributions in each neuron of an LSTM-based model. Our LSTM-based en-

ergy demand forecasting model was used to predict hourly, daily and weekly energy demand effectively on two different datasets for five different households in terms of all evaluation metrics.

The major goal of this study was to explain the predictions in such a way that users can have a clear understanding of why a particular decision has been predicted. Our framework applied DeepLIFT to approximate the SHAP values to generate easy-to-understand explanations. These explanations generation technique combining DeepLIFT and SHAP can be applied to interpret the predictions for any deep learning-based forecasting models. The explanations can highlight both the time or season and the impact of different attributes (features) for a particular prediction at the same time. Based on our introduced evaluation metric named contribution monotonicity coefficient, the generated explanations achieved high efficiency and the relationship with original contributions of different appliances toward the total consumption is monotonous.

We also observed that the explanations for household energy forecasting can identify the impacts of appliances for corresponding energy consumption activities that are aligned and correlated with the findings of the previous study [302]. With these explanations, users might be more aware of and think of optimizing their energy consumption practice by considering the most responsible factors for their upcoming energy consumption demand. The predicted future energy demand and the explanation can help the users with energy consumption literacy [268, 270] and the policymaker can think of adopting our method in dynamic pricing and energy policy optimization.

6.6 Future Direction

In the pursuit of creating functional user interfaces tailored to smart home users, we will adopt a user-centric design methodology, akin to the approach advocated by Rikke et al. [140]. Our current trajectory involves the development of a prototype for our proposed system, which aims to offer transparent insights into energy demand prediction and forecasting. Our underlying assumption is that by allowing smart home users to interact with our prototype in their daily lives, we can glean insights into both the domain and the technology. This interaction will provide them the opportunity to articulate the types of explanations they consider vital and valuable. This approach is especially significant in light of existing systems, often geared towards developers and AI experts, which may exhibit certain limitations. In this regard, we have outlined a set of inquiries enumerated below, which we intend to pose as we construct our explainable prediction system with a strong emphasis on human-centered design.

1. Are these explanations helpful for you in understanding the decision-making process?
2. What open or further questions would you like to have answered, if any?
3. Do you find the presented user interface useful for engaging with the presented explanations?
4. What problems or areas for improvement would you see in this respect, if any?
5. Thinking aloud, would you please walk us through the explanation interface, reflecting on a particular prediction that is presented there?

We believe that through a user-centered prototyping approach with different kinds of explanation visualizations, we can learn more about the specific user needs in the energy domain, and elicit requirements and insights towards building a collaborative, human-centered explainable energy demand forecasting system. Hence, the system will increase transparency, fairness, and accountability to end-users.

Acknowledgment

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 955422.

7 Improved Thermal Comfort Model Leveraging Conditional Tabular GAN Focusing on Feature Selection

The content of this chapter has been presented in the 1st ACM International Workshop on Big Data and Machine Learning for Smart Buildings and Cities (ACM Balances) co-located with the Thirteenth ACM International Conference on Future Energy Systems 2022, Boston, Massachusetts, USA. The paper has been published in the proceedings of ACM BuildSys2022 as a short paper. The extended version later has been published as a full paper in the *IEEE Access*, published by IEEE in 2024. The information of both papers is given as follows:

Infrmation of Article 1: Md Shajalal, Milad Bohlouli, Hari Prasanna Das, Alexander Boden, and Gunnar Stevens 2022. Focus on What Matters: Improved Feature Selection Techniques for Personal Thermal Comfort Modelling, In *Proceedings of the 9th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (BuildSys '22)*. Association for Computing Machinery, New York, NY, USA, 496 – 499. https://doi.org/10.1145/3563357.3567406

Information of Article 2: © 2024 IEEE. Reprinted, with permission, from Md Shajalal, Milad Bohlouli, Hari Prasanna Das, Alexander Boden, and Gunnar Stevens. 2024. Improved Thermal Comfort Model Leveraging Conditional Tabular GAN Focusing on Feature Selection. *IEEE Access*, IEEE vol. 12, 30039–30053. https://doi.org/10.1109/ACCESS.2024.3366453

Abstract

The indoor thermal comfort in both homes and workplaces significantly influences the health and productivity of inhabitants. The heating system, controlled by Artificial Intelligence (AI), can automatically calibrate the indoor thermal condition by analyzing various physiological and environmental variables. To ensure a comfortable indoor environment, smart home systems can adjust parameters related to thermal comfort based on accurate predictions of inhabitants' preferences. Modeling personal thermal comfort preferences poses two significant challenges: the inadequacy of data and its high dimensionality. An adequate amount of data is a prerequisite for training efficient machine learning (ML) models. Additionally, high-dimensional data tends to contain multiple irrelevant and noisy features, which might hinder ML models' performance. To address these challenges, we propose a framework for predicting personal thermal comfort preferences, combining the conditional tabular generative adversarial network (CTGAN) with multiple feature selection techniques. We first address the data inadequacy challenge by applying CTGAN to generate synthetic data samples, incorporating challenges associated with multimodal distributions and categorical features. Then, multiple feature selection techniques are employed to identify the best possible sets of features. Experimental results based on a wide range of settings on a standard dataset demonstrated state-of-the-art performance in predicting personal thermal comfort preferences. The results also indicated that ML models trained on synthetic data achieved significantly better performance than models trained on real data. Overall, our method, combining CTGAN and feature selection techniques, outperformed existing known related work in thermal comfort prediction in terms of multiple evaluation metrics, including area under the curve (AUC), Cohen's Kappa, and accuracy. Additionally, we presented a global, model-agnostic explanation of the thermal preference prediction

system, providing an avenue for thermal comfort experiment designers to consciously select the data to be collected.

Keywords

Personal Thermal Comfort, Generative Adversarial Network, Feature Selection, Machine Learning, Data Inadequacy

7.1 Introduction

Occupants' well-being, health, and productivity significantly depend on thermal comfort both at home and in the workplace [281, 191, 49, 190, 77, 96]. A notable portion of the total energy consumption is attributed to the HVAC (heating, ventilation, and air conditioning) system, accounting for nearly half of the overall energy use in corporate and residential buildings [77]. Additionally, these buildings contribute to almost 40% of CO_2 gas emissions [281, 77]. The advancements in sensor technology over the last two decades have played a crucial role in shaping the concept of smart home systems to reality, empowering inhabitants to control and monitor the indoor environment within their homes and workplaces [191, 49, 190, 77, 96, 278]. Environmental parameters related to thermal comfort, such as temperature and humidity, can be adjusted using multiple machine learning-based systems with human-in-the-loop interaction [281].

In general, artificial intelligence (AI)-based techniques can be applied to have energy-efficient and comfortable indoor environment inside buildings [201, 220]. It is also evident that researchers often leveraged AI-enabled techniques for energy aware and comfortable built environment. However, the primary objective is to save energy and decrease the carbon-di-oxide footprints. The notable smart home energy-aware applications that generally applied deep learning (DL) and ML models can be

energy demand forecasting, adjusting indoor environment by predicting thermal comfort preferences, etc [281]. However, in this work we focus on personal thermal comfort preference prediction. Recently, there has been a considerable attention in applying ML models for thermal comfort preference prediction tasks [220, 1, 327, 60, 59, 64, 299, 106, 94, 92, 114, 61, 281].

Generally, the task of personal thermal comfort preference prediction can be classified into two different categories, global and personal. In global thermal comfort (GTC) preference prediction task, the model tries to predict the overall thermal comfort preference in the rooms/zones. On the other hand, since the thermal comfort of different person varied widely, personal thermal comfort (PTC) preference prediction refers to identifying an occupant's individual thermal comfort [281]. Based on the preference prediction system's output, smart home systems can control and adjust the environment to provide pleasant and comfortable living space.

In most of the existing studies [191, 1, 92, 299], authors applied their predictive models to high dimensional features to capture relation between the data and occupants' thermal preference. There are two major challenges associated with modeling the personal thermal comfort preference prediction: one is the high dimensionality of the data including environmental and physiological features, and other one is the lake of adequate amount of data samples to train efficient predictive ML model. This is expected that the data to predict thermal comfort preference will be high dimensional, since it considers every possible attributes that are related to the occupants indoor HVAC comfort. On the other hand, the data collection from real subjects with right annotation procedure is very time consuming and costly.

Generally, ML models needs adequate data to train and this is a prime requirement in any predictive models. To mitigate the data availabil-

ity problem, an effective synthetic data generation technique addressing associated challenges can be a game changer. The high-dimensionality might be the curse in modeling indoor thermal comfort preference. Because there might have some features that are not relevant and can even downgrade the performance of the predictive model. Hence, identifying the possible relevant set of features is a prerequisite of the system with high-dimensional data.

In this research, we propose a new indoor thermal comfort preference prediction system by addressing the above-mentioned challenges by incorporating CTGAN and multiple feature selection techniques. First, we address the data inadequacy challenge by employing one of the most successful synthetic data generation techniques that incorporate the multi-modal distribution in the numeric features with mode-specific normalization technique. In addition with the data adequacy problem, datasets related to the PTC preference are generally imbalanced, which might make the performance biased towards the majority class samples. By incorporating CTGAN, we address also the data imbalance problem by synthetically generating the data for minority class samples.

The best set of relevant features generally provides high performance in predictive modeling in case of high-dimensional datasets. Before applying feature selection techniques, we conducted experiments to determine whether highly correlated features exist in the PTC dataset. Our hypothesis for this experiment was that if we found more correlated features related to PTC preference prediction, we could then make use of feature selection techniques to filter out irrelevant, noisy, and redundant features. To achieve this, we carried out experiments on a PTC preference prediction dataset [191]. The correlation among the 82 features, based on Pearson's correlation coefficient analysis, is illustrated in the heatmap representation in Fig. 25.

We observed that more than 20 features out of the 82 different phys-

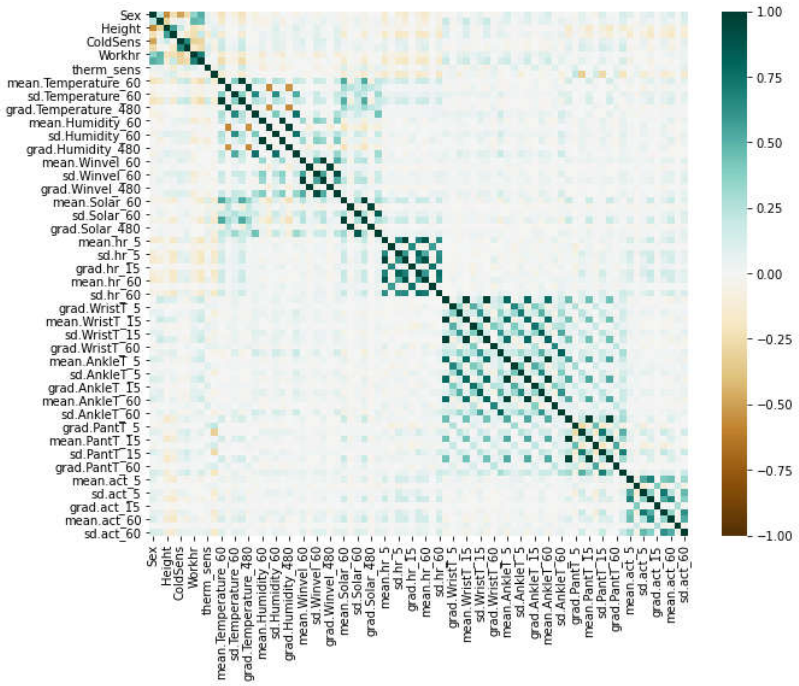

Figure 25: Heatmap depicting correlation coefficient among different features

iological, environmental, and weather features are highly correlated ($\rho \geq 0.80$) [281]. With these findings, we developed the idea to apply various feature selection techniques to filter out irrelevant features. Our previous paper, based on the preliminary findings on the effect of applying feature selections, is published in the proceedings of the ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (ACM BuildSys 2022) [281]. Inspired by the impressive preliminary findings, this extended research applied multiple feature selection techniques and introduced CTGAN to generate synthetic data samples for effectively training ML models.

We conducted experiments by training six different ML models on a standard PTC prediction dataset by generated synthetic data samples employing CTGAN for thermal comfort prediction. We leveraged the best-selected feature set, applying multiple feature selection techniques, for

training the ML models. Since the dataset was imbalanced, we carefully utilized multiple evaluation metrics, including *Cohen's Kappa* and *Area Under the Curve (AUC)*, along with *accuracy*, that can measure the prediction performance a ML model for imbalance data distribution.

The experimental results, encompassing a wide range of settings, demonstrated the superiority of the proposed method with synthetically generated data focusing on feature selection techniques. The performance of predictive models trained on synthetic data significantly outperformed the baseline, as well as models trained on real data with feature selections. Compared to known related works, our methods also achieved much higher performance in terms of all evaluation metrics (AUC, Kappa, and Accuracy). The contributions of this research are summarized as follows:

- We introduced CTGAN, a synthetic data generation technique, to address the problem of data inadequacy by generating new personal thermal comfort data for individuals.
- We employed multiple feature selection techniques to identify the best possible set of relevant features for effectively modeling personal thermal comfort preference prediction.
- The experimental results demonstrated the superiority of our framework in modeling thermal comfort preference prediction. We achieved significantly higher performance after applying feature selection techniques and CTGAN, a synthetic data generation technique. The combination of both techniques showed a significant improvement in performance compared to known related methods.

In the remainder of the paper, we present state-of-the-art on personal thermal comfort preference prediction in section 7.2. We then present our proposed thermal comfort modeling framework combining on CTGAN and feature selection techniques in section 7.3. In section 7.4, we

discuss about the dataset, evaluation metrics, experimental design and the findings by demonstrating results for wide range of experiments. Finally, we conclude our findings with future research direction in section 7.5.

7.2 Literature Review

The prior works on modeling PTC preferences are associated with the experiments on the data collected from living labs [152, 26, 139, 177, 179, 180, 289]. Generally, a large number of features are included, often the amount of data samples available after cleaning the data of any missing value in the features decreases [152, 177, 289], especially for experiments including physiological metrics for an occupant. Consequently, classical ML models often have more predictive power compared to deep learning based models [100, 325, 198, 99]. Therefore, unlike the latter models, which automate the feature engineering by learning from the data, with the classical models, feature engineering plays a key role in the predictive power.

However, most of the datasets for PTC preference prediction task are small in sample size and the sample distributions among different classes are quite imbalanced. Since ML models need adequate data for learning the pattern from the samples, it is challenging to train models with small dataset. In addition, collecting big dataset from the participants in a living lab setting is quite time consuming and costly. Therefore, synthetically generating data samples on the available data has got considerable attention in thermal comfort modeling research in recent time [237, 334, 78, 329, 328].

To address the data inadequacy challenges in thermal comfort preference prediction task, several methods has been proposed that generate synthetic data [237, 334, 78]. Synthetic data generation techniques

are also often applied for balancing the data [277, 287]. Quintana et al. [237] employed conditional generative adversarial networks to generate synthetic data samples for minority class to address class imbalance problem. Similarly, conditional Wasserstein GAN has been applied by Yoshikawa et el. [334] for balancing the thermal comfort preference prediction dataset. Das et al. [78] also applied basic GAN architecture for the same purpose.

Evidently, the inclusion of redundant features degrades the performance of the ML model [7]. Feature selection has also been of interest of similar fields, e.g. in occupancy prediction where the objective is to predict the occupancy count of the rooms in a building, adaptive lasso feature selection has been used to select the most relevant features [316]. Similarly in [95] authors use genetic algorithms for feature selection.

Feature selection techniques based on manual observation that evaluate the best combination by the prediction performance of the trained model. For instance in [152], authors tried out the various combinations of the input features and concluded that skin temperature and heating settings are the best predictors for thermal comfort. In a similar study [289], authors examined the skin temperature at 6 points on the body and evaluated the various combinations, besides also proposing a new feature representative of the body's average temperature based on Ramanathan's formula as a combined feature [289]. In another study [194], authors defined 3 feature sets and evaluated their predictive power via precision and recall. The above-mentioned approaches, despite being the most prevalent approaches adopted in the literature, require significant background knowledge regarding the features and significant manual labor. Also, the manual approach might not necessarily result in the best combination, especially for datasets with high dimensions.

Another selection method is related to the prior knowledge from thermal

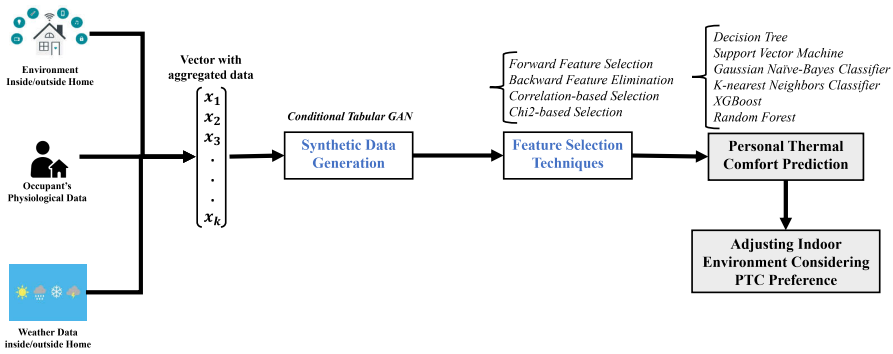

Figure 26: A high-level building block of the proposed PTC preference prediction with synthetic data using CTGAN focusing on feature selection

comfort domain and literature. For example, in [343], authors have defined new features based on the ASHRAE standard [320] which would be representative for model structure and heat balance of the body. In another study [115], authors derived new features based on the polynomial basis function to capture the relation between the environmental features and thermal perception. Although these groups of studies may introduce new crucial features that may promote the predictive power, they still have to employ a selection process to omit the less productive features. Similar to the first group of the studies, this approach also requires prior knowledge while introducing new features.

One of the most classical ways is to apply feature selection techniques to find the relation between the input variables and the target variable. For example, in the study conducted in [65], authors tried to find the relation between the thermal sensation and the air quality via multilinear regression and hypotheses testing. However, in their study they do not differentiate between heating and cooling. Similarly, in [114], authors used Lasso feature selection to select among the features that collected from an experimental study. Additionally, in the study done

in [191], authors employed Pearson correlation coefficient among the features and the target to measure the importance of the introduced features.

In this research, we applied conditional tabular GAN [329, 328] inspired by the success achieving new benchmark. CTGAN can solve the multimodal distributions in the numeric data points by incorporating mode-specific normalization technique and address challenge associated with categorical features applying variational Gaussian-Mixture model [329, 328]. Adequate training data, complemented by synthetic data generated through CTGAN, ensures the effective training of ML models. Furthermore, this can address the issue of data imbalance in the training set, ensuring that the model does not exhibit bias towards any majority class. On the other hand, we incorporated multiple feature selection techniques to identify the best set of features that help the ML models to achieve higher performance in thermal comfort preference prediction.

7.3 Methodology

The overview of our proposed PTC preference prediction framework is illustrated in Fig. 26. We first generate synthetic data applying CTGAN and then apply four different feature selection techniques to identify the best possible sets of features. With the selected features, we employed six different ML classification models to predict occupants' thermal comfort preference.

7.3.1 Synthetic Data Generation with CTGAN

A significant challenge in developing predictive models for PTC preferences arises from the scarcity of adequate data. The dataset [191] utilized in our experimental work, as described in greater detail in Sec-

tion 7.4.1 was sourced from a group of 14 individuals. These subjects participated in data collection activities that involved the annotation of their thermal comfort preferences while residing in living laboratories located in Berkeley and San Francisco.

It is worth noting that acquiring large datasets specially for training of state-of-the-art ML models for PTC preference prediction is expected to be very expensive and time consuming. Consequently, there is potential value in generating synthetic data through the application of robust data generation techniques. This approach is inspired by the remarkable achievements of generative adversarial networks (GAN) when applied to tabular data, as evidenced by prior works [328, 329]. Specifically, we explore the application of conditional tabular GAN (CTGAN), which offers particular advantages in addressing challenges related to mixed data types and multi-modal distributions when generating synthetic tabular data.

Unlike other GAN-based methods including WGAN [17] and WGAN-GP [116], CTGAN can capture the heterogeneity of the real-world data [328, 329]. To handle mixed data in creating synthetic data, CTGAN developed a full workflow from data preprocessing to modifying GAN architecture. The major challenge that CTGAN solved is non-Gaussian multimodal distribution by introducing a mode-specific normalization technique. It handles this problem by following multimodal distributions. By applying a variational Gaussian mixture model (VGM), it can represent each continuous real-valued feature in a one-hot vector that indicates the sampled mode and the normalized value [328, 329]. To tackle challenges posed by categorical features, CTGAN introduced the sparsity of one-hot-encoded vectors in real-valued data with probability distributions [328, 329]. Further, it introduced a conditional data generator that gets ride of the challenges posed by multimodal and imbalanced data distributions. The detail description of CTGAN can be

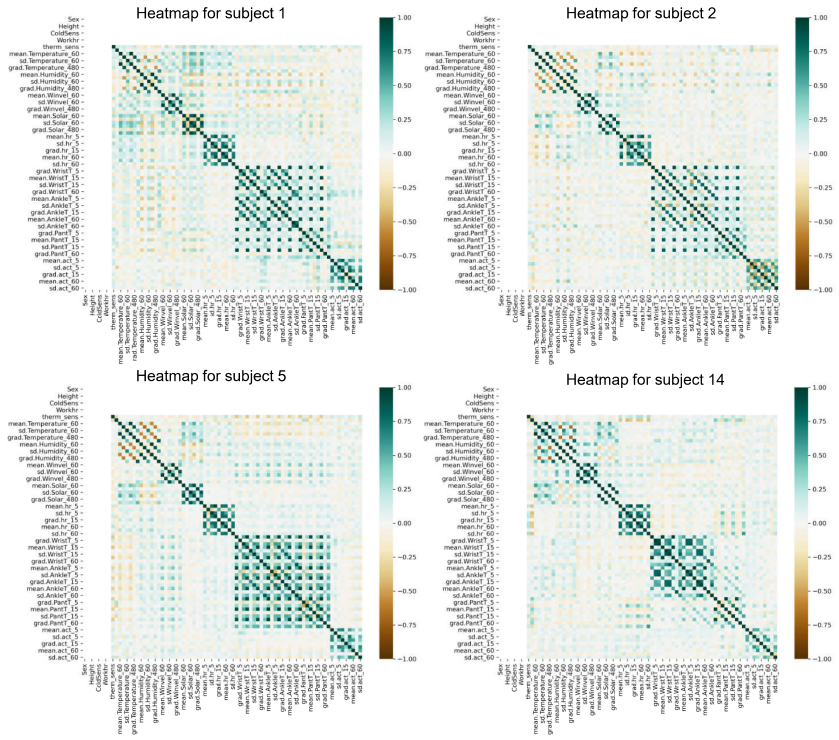

Figure 27: Heatmaps for four different subjects highlighting the correlation among different features

found in [328, 329].

7.3.2 Feature Selection Techniques

Classifiers are often misled by redundant, correlated and noisy features. In case of a high dimensional dataset, selecting best features' set before applying classifier would be a better approach for modeling thermal comfort. In this section, we will outline our visual exploration of feature redundancy. This exploration indicates the usefulness of feature selection techniques to filter out less relevant features. Next we introduce four feature selection techniques in our study to filter out redundant and correlated features to improve the performance of PTC prediction model.

Visual exploration: We explore the correlation among different features

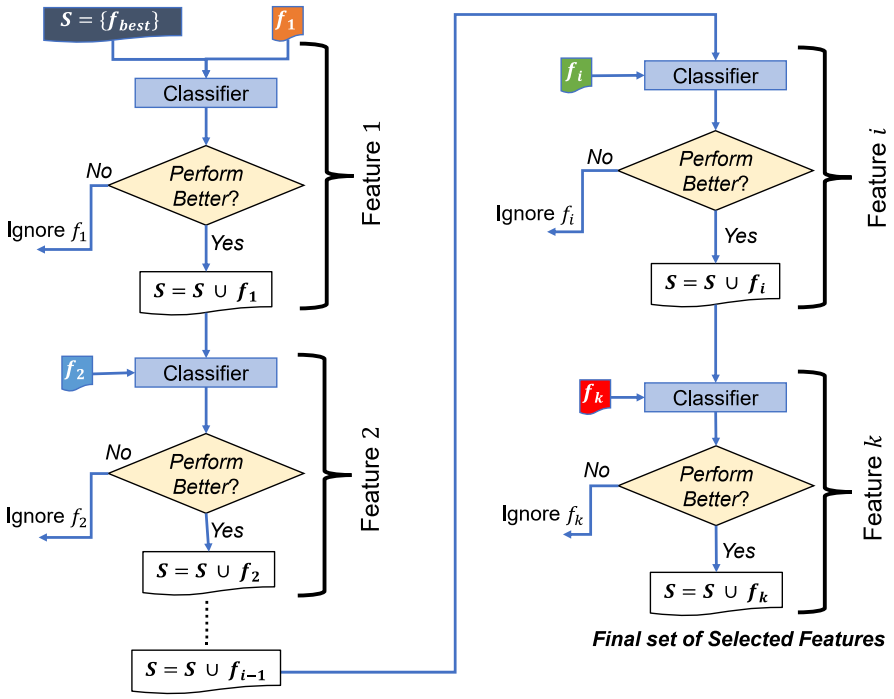

Figure 28: The workflow of the forward feature selection (FFS) technique (Figure created based on [281])

across occupants numerically as well as visually to detect patterns in them. Heatmaps in Fig. 27, show the feature correlation for four different occupants. We can see that there are a substantial number of features that are correlated to each other. However, we can also see that the correlation coefficients among different features are quite different among different occupants. These observations and findings illustrate why PTC prediction is a challenging task. The detailed view also shows that there are some common patterns and correlations among some features across the occupants. With this preliminary analysis, we hypothesize that the elimination of these correlated and redundant features might improve the PTC prediction model.

Correlation-based feature selection: Since one of our primary objectives is to filter out irrelevant features before applying classifiers, we conducted a correlation analysis across all features. In correlation-based feature selection, we consider a feature as redundant if it has a high

correlation coefficient ($\rho \geq 0.80$) and remove it from the list [281]. We compute the Pearson correlation coefficient between two features as follows:

$$\rho(f_a, f_b) = \frac{\sum_{i=1}^{n}(f_{a_i} - \bar{f}_a)(f_{b_i} - \bar{f}_b)}{\sqrt{\sum_{i=1}^{n}(f_{a_i} - \bar{f}_a)^2 \cdot \sum_{i=1}^{n}(f_{b_i} - \bar{f}_b)^2}} \tag{7.1}$$

where f_a and f_b are two features from the list of features $F = \{f_1, f_2, f_3, \ldots, f_k\}$. The average feature values for two different features f_a and f_b are denoted by $\bar{f}_a$ and $\bar{f}_b$, respectively [281].

Chi-Square test-based feature selection: The target of this technique is to select those features, which have higher dependency with the response. In statistics, Chi-Square test is a prominent technique applied to test the independence of two different events. In this research, however, we employed Chi-Square test as a tool to select the best set of features. Given two different variables, Chi-square test computes how the expected count E deviates from the observed count O for those two variables. The computation is done as $\mathscr{X}^2 = \sum \frac{(O_i - E_i)^2}{E_i}$. We employ this test to determine the relationship between specific features and the labeled response. Chi-Square will return a smaller value when two features are independent. In other words, the observed count is close to the expected count. Hence, the higher the Chi-Square value the feature is more dependent on the response and that feature should be selected for training the model. We applied iterative approach to have Chi-Square test for each feature and selected the best features' set. To select an optimal value for the number of selected features, we make use of a grid search. The details on parameter tuning is presented in section 7.4.4

Supervised forward feature selection: We applied supervised forward feature selection (FFS), which enables the selection of more insightful features through a greedy iterative selection procedure. We present the FFS procedure in Fig. 28 [281]. This approach first utilizes every feature individually and applies a baseline classifier to predict occupants' personal thermal comfort. By comparing the performance of all individual

features, it selects the best-performing one as f_{best}. The first selected feature f_{best} is then added to the selected feature set S. Subsequently, FFS then combines the remaining features f_i one at a time to the selected feature set S and applies the classifier separately [281]. Considering the performance of the classifier, FFS selects the feature f_i if the combination achieved better performance than the previous classifier with the already selected feature set S. This greedy approach continues for the rest of the features. Applying this FFS approach allows us to select the best set of features that are effective in predicting occupants' PTC preferences efficiently [281].

Supervised backward feature elimination: The working principle of backward feature elimination (BFE) is the opposite of forward feature selection. Unlike the forward feature selection approach, it first applies all the features feeding to a classifier to model PTC and then computes the classification performance. After computing the performance, it iteratively discards one feature at a time and checks whether the performance of the model increases or decreases without that feature. If the performance decreases, then it hypothesizes that the feature has an important role in modeling PTC. Consequently, it includes that feature in the list of important features. In the opposite case, it ignores the feature and discards it as less relevant.

7.4 Experiments

This section presents the wide range of experimental setups and performance evaluations of our proposed methods that validate the efficiency in modeling personal thermal comfort in terms of multiple evaluation metrics on a PTC dataset.

7.4.1 Dataset

We conducted experiments on a PTC preference prediction dataset collected by Liu et al. [191]. The dataset was collected and annotated by 14 different subjects living in the areas of Berkeley and San Francisco. During the study, the authors measured the skin temperature from different parts of the subjects' bodies and the surrounding room temperature where the subjects were present. Additionally, they also measured the activity and heart rate of the subjects using accelerometers and polar sensors, respectively. The experiments spanned 14 days, with each subject expected to provide their thermal comfort preference 12 times a day, categorized as "Cooler," "Warmer," or "No Change." Out of the collected 3848 samples, the distribution across different classes is quite imbalanced. Fig. 29 illustrates the percentage of samples across different classes, showing that 68.5% of samples are for "No Change," while 16.5% and 15% are for "Cooler" and "Warmer," respectively.

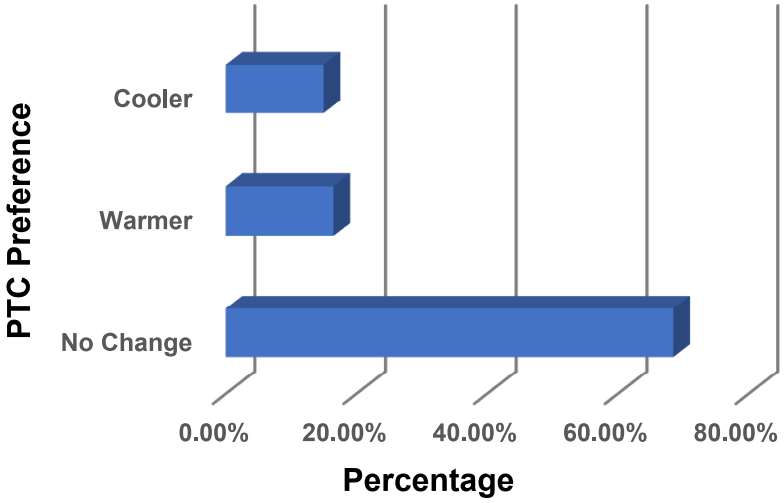

Figure 29: Distribution of samples over PTC preferences

7.4.2 Data Pre-Processing

The values of the features in the dataset vary widely in terms of their units and ranges. In addition, there are some missing values which we tackled by applying median values. After that we applied min-max normalization [135] to map each variables' values to a certain range [0,1]. We also grouped the dataset based on individual occupant and analyzed the features for individual occupants.

7.4.3 Evaluation Metrics

In the assessment of any ML model, it is crucial to consider the evaluation metrics with respect to the characteristics of the dataset. According to the literature on PTC models [190], Accuracy, AUC (Area Under Curve), and Cohen's kappa are the three widely used evaluation metrics.

The validity and usefulness of evaluation metrics also depends on the specific domain and characteristics of the datasets. In our case, it is essential that the evaluation metric are sensitive to the class-imbalance. For instance, using accuracy alone will be problematic, since this metric will not reflect the class-wise prediction performance. Considering the imbalance distribution in the personal thermal comfort datasets, a model classifying all the samples as "no change" with an 80% share in the original dataset, would result in an 80% accuracy, which does not necessarily suggest the strength of the classification.

In the following formulations, True Positive (TP) is an outcome where the model correctly predicts the positive class, True Negative (TN) is an outcome where the model correctly predicts the negative class, False Positive (FP) is an outcome where the model incorrectly predicts the positive class, False Negative (FN) is an outcome where the model incorrectly predicts the negative class.

Accuracy: As shown in equation 7.2, *Accuracy* only requires the class

labels for evaluation and does not examine the separability strength of the model. Nonetheless, it has also been reported to be compared with the previous studies.

$$accuracy_{class1} = \frac{TP+TN}{TP+TN+FP+FN} \tag{7.2}$$

Area under the curve (AUC): In contrast to Accuracy, AUC [138] also considers how well the predicted classes are distinguished, by taking the prediction probabilities of each class into account. Technically, AUC is the area under the curve of the ROC (Receiver Operative Characteristics) which is the representation of TPR (True Positive Rate) with respect to the FPR (False Positive Rate), defined in equations 7.3 and 7.4, respectively when the decision boundary is moved through the data points.

$$TPR = \frac{TP}{TP+FP} \tag{7.3}$$

$$FPR = \frac{FP}{TN+FP} \tag{7.4}$$

Fundamentally, this metric is proposed for binary classification problem, however, in order to apply it to the multiclass cases, the One vs Rest approach has been used.

Cohen's Kappa: Cohen's Kappa[138] is often an under-utilized, but quite useful metric, which also considers the prediction probabilities of each class. It can be defined as follows:

$$kappa = \frac{P_0 - P_c}{1 - P_c} \tag{7.5}$$

$$P_0 = \frac{TP+TN}{TP+TN+FP+FN} \tag{7.6}$$

$$P_c = P(\text{“}Positive\ Classified\text{”}) + P(\text{“}Negative\ Classified\text{”}) \tag{7.7}$$

$$P(\text{“}Positive\ Classified\text{”}) = \frac{TP+FP}{TP+TN+FP+FN} \tag{7.8}$$

$$P(\text{“}Negative\ Classified\text{”}) = \frac{TN+FN}{TP+TN+FP+FN} \tag{7.9}$$

Cohen's kappa considers the quantity of the classes. More precisely, it consider the probability of classes being changed, defined as P_C (the observed agreement) and P_0 (the expected agreement) . It varies from 0 to 1, with 0 being a random classification.

7.4.4 Feature Selection

We applied the four feature selection techniques described in section 7.3. The number of selected features differ among selection techniques.

Chi-Square technique's parameter tuning: The Chi-Square-based feature selection technique requires to tune the parameter k, the number of selected features. We applied a grid search to identify the optimal number of selected features and evaluated the performance. The experimental results in terms of AUC are illustrated in Fig. 30. The figure concludes that the optimal number of features (highest AUC) should be $k = 17$ and the performance with those selected features are on the y-axis. Therefore, we select $k = 17$ and apply this parameter value in our chi-square-based feature selection.

Result of feature selection techniques: None of the introduced feature selection techniques need parameter tuning except the Chi-Square test. For correlation-based selection, we make use of $\rho >= 0.80$ (Eq. 7.1) as highly correlated features. In turn, we could apply other techniques straight forward. Table 12 presents the detailed results of our feature selection techniques with the number of selected features. The four fea-

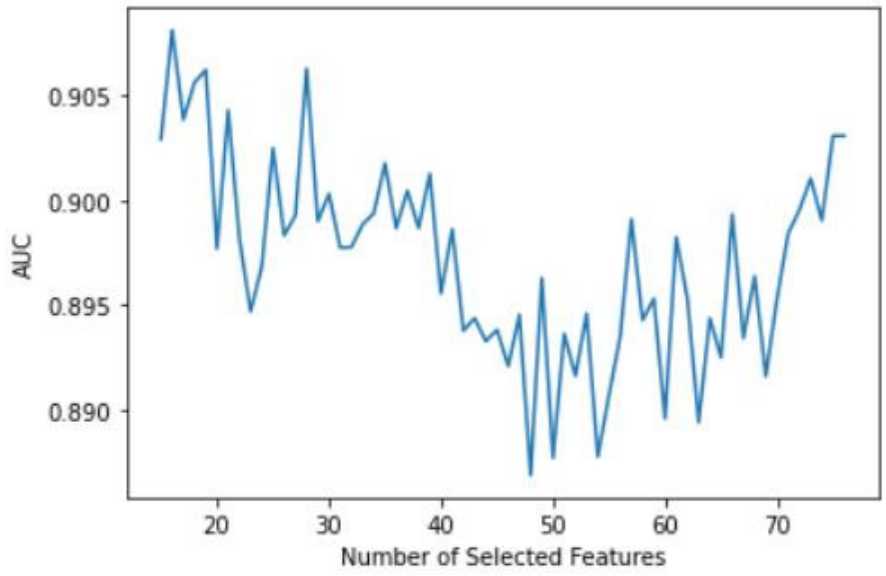

Figure 30: Tuning the parameter k, number of selected features in Chi-Square feature selection using grid search

Table 12: Results of applying feature selection techniques and notable selected features

Features	**Chi-Square**	**Correlation**	**FFS**	**BFE**
# of features	17	59	27	32
Age	x	x	x	x
Height	x	x	x	x
Therm_sens	x	x	x	x
Temperature	x	x	x	x
Workhr	x	x	x	x
Hear rate	x	x	x	x
Sex	-	x	-	x
ColdSens	-	x	x	x
ColdExp	-	x	-	x
Coffeeintake	-	x	-	x
Location	x	x	x	x
Humidity	-	x	-	x
WristT	x	x	-	x
Winvel	-	x	x	-
AnkleT	-	x	x	x
PantT	-	x	x	x
Solar	x	x	x	x
Act	x	-	x	x

Table 13: Performance of applying feature selection techniques in different ML models trained on real data in global thermal comfort prediction. The best results for each feature selection techniques are in **bold**. The blue-colored values indicate the best performance among all experimental settings.

Feature Selection	Model	Kappa	Accuracy	AUC
Chi2-based Selection	DT	0.6331	0.8374	0.9229
	SVM	0.5977	0.8270	0.9029
	KNN	0.4966	0.7906	0.8574
	BNB	0.4644	0.7464	0.8199
	XGB	**0.6585**	0.8491	**0.9401**
	RF	0.6567	**0.8517**	0.9378
Correlation-based Selection [281]	DT	0.6551	0.8348	0.8964
	SVM	0.5054	0.8023	0.9101
	KNN	0.2830	0.7308	0.7802
	GNB	0.2825	0.5825	0.7267
	XGB	**0.6830**	**0.8595**	**0.9370**
	RF	0.5417	0.8127	0.9167
Forward Feature Selection [281]	DT	0.6471	0.8426	0.8852
	SVM	0.5441	0.8127	0.9167
	KNN	0.3939	0.7646	0.8184
	GNB	0.4627	0.7256	0.7923 5
	XGB	**0.6782**	**0.8569**	0.9362
	RF	0.6109	0.8374	**0.9380**
Backward Feature Elimination	DT	0.6709	0.8426	0.9007
	SVM	0.5474	0.8140	0.9227
	KNN	0.3733	0.7542	0.8029
	GNB	0.4291	0.7113	0.7740
	XGB	**0.6717**	**0.8556**	**0.9438**
	RF	0.5895	0.8296	0.9375

ture selection techniques chi-square based, correlation based, forward feature selection and backward feature elimination are denoted as Chi-Square, Correlation, FFS and BFE, respectively. As outlined above, in the case of Chi-Square-based feature selection technique, the number of selected features was a result of our parameter tuning. We can see in Table 12, some features are common in all selected feature sets. From 82 different features, we enlisted here the most notable 19 features. Of these 19 features, some features have different varieties i.e., the temperature has different varieties such as mean, gradient, and standard

deviation of different time slots.

Similar to temperature, skin temperature on the wrist and ankle, wind speed, and wrist acceleration also have some varieties. Here in this list, we checked if any one of the varieties is selected by the feature selection technique. However, we can see that *age, height, thermal sensitivity, temperature, working hour, heart rate, weight, and subject location* are the common features that all the selection techniques selected as relevant. Other than that, the skin temperature of the wrist, ankle, and body proximity temperature are important features considered by the three selection techniques. Similarly thermal sensitivity, cold sensitivity and cold extremity experience also came out to be important in modeling PTC preference.

7.4.5 Experimental Setting

We first applied feature selection techniques and trained the ML models on original data samples collected from 14 different subjects. For all of these feature sets selected based on selection criteria, we carried out a range of experiments to validate the performance of our introduced feature selection techniques. At first, we applied our methods on the whole dataset combining samples from all 14 subjects. The primary intuition was to observe how our feature selection methods perform on overall thermal comfort dataset. In other words, evaluating our methods on the global thermal comfort (GTC) preference in buildings. Then we applied the similar experimental setting to train ML models on synthetically generated data by CTGAN with the selected relevant features sets.

Finally, with the selected features leveraging four different feature selection techniques, we applied six different classical classifiers to model the thermal comfort of the occupants. Then, we applied two best per-

Table 14: Performance of applying feature selection techniques in different ML models trained on synthetic data with CTGAN in global thermal comfort prediction. The best results for each feature selection techniques are in **bold**. The blue-colored values indicate the best performance among all experimental settings.

Feature Selection	Model	Kappa	Accuracy	AUC
Chi2-based Selection	DT	0.6400	0.7629	0.8857
	SVM	0.6944	0.7983	0.9245
	KNN	0.5410	0.7020	0.8716
	GNB	0.5746	0.7206	0.8779
	XGB	**0.8265**	**0.8854**	**0.9734**
	RF	0.8054	0.8713	0.9638
Correlation-based Selection	DT	0.5986	0.7366	0.8760
	SVM	0.7449	0.8320	0.9404
	KNN	0.5469	0.7067	0.8702
	GNB	0.5827	0.7261	0.8671
	XGB	**0.8483**	**0.8998**	**0.9786**
	RF	0.8008	0.8686	0.9602
Forward Feature Selection	DT	0.6537	0.7687	0.8836
	SVM	0.7885	0.8603	0.9592
	KNN	0.5265	0.6957	0.8859
	GNB	0.5676	0.7170	0.8715
	XGB	**0.8872**	**0.9253**	**0.9890**
	RF	0.8282	0.8860	0.9765
Backward Feature Elimination	DT	0.6673	0.6673	0.8973
	SVM	0.7983	0.7983	0.9606
	KNN	0.4954	0.6769	0.8730
	GNB	0.6164	0.7485	0.8954
	XGB	**0.8802**	**0.9208**	**0.9885**
	RF	0.8455	0.8979	0.9788

forming models to observe the performance on PTC preference prediction for individual subject. By doing so, we applied our method on 14 different subjects' collected samples separately to model PTC. The remainder of the section presents the experimental results for modeling global and personal thermal comfort and performance comparison with prior works. We repeat all experiments for GTC and PTC with training the models with our generated synthetic data by conditional tabular generative adversarial networks for modeling thermal comfort preference prediction.

7.4.6 Global Thermal Comfort Prediction Performance

We designed the experiments in a way to visualize the performance of the feature selection techniques in modeling thermal comfort preference with the trained models on real data. As we noted earlier that the data might not be adequate to train ML models, we employed CTGAN to generate quality synthetic data and combine it with real data for training the model with feature selection. Therefore, we first illustrate the performance of different feature selection techniques on real data and then we present the results on applying CTGAN for synthetic data generation.

Performance on real data: The performance of six different classifiers trained on real data samples of the global thermal comfort dataset with our introduced feature selection techniques is summarized in Table 13. With the selected features for each feature selection technique, we applied six different classifiers: decision tree (DT), support vector machine (SVM), K-nearest neighbor (KNN), Gaussian Naive Bayes (GNB), XGBoost (XGB), and random forest (RF). This results in a total of 24 experimental setups.

Table 13 shows that XGBoost with the correlation-based feature selection technique achieved better performance among all experimental settings (highlighted in blue) in terms of Cohen's Kappa and Accuracy. However, in terms of AUC, the XGBoost model with the backward feature elimination technique obtained the best performance. Among the six different ML models, the XGBoost model is consistently better across all feature selection techniques, except for Forward feature selection and Chi-square-based selection techniques. In both cases, random forest (RF) acquired the highest accuracy for the Chi-square-based selection procedure and the best AUC for the forward feature selection technique. However, it is also observed that the performance difference among correlation-based, forward feature selection, and backward fea-

ture elimination is not significant. We can broadly say that among all six different classifiers, XGBoost and random forest are the two best-performing models across all feature selection techniques.

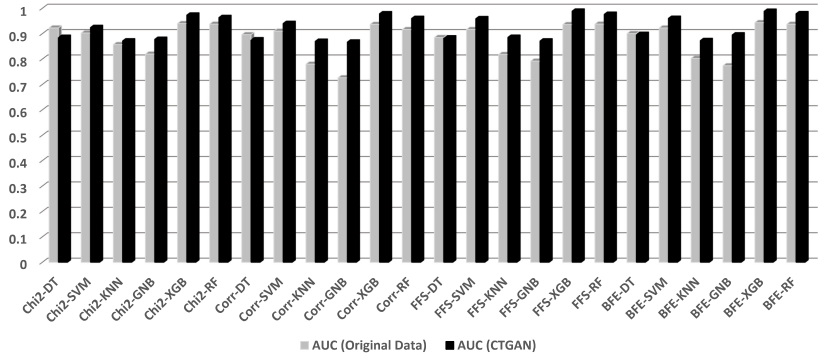

Figure 31: The performance comparison of all experimental settings between the models trained on original data and synthetically generated data by CTGAN, respectively.

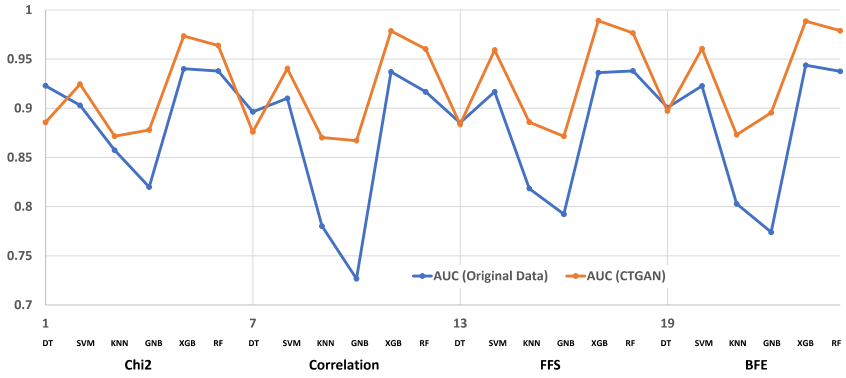

Figure 32: The performance comparison of all experimental settings between the models trained on original data and synthetically generated data by CTGAN, respectively.

Performance on Synthetic Data: The performance of introduced feature selection techniques with the synthetically generated data employing the prominent conditional tabular GAN is presented in Table 14. The result shows clearly that the performance has been improved significantly for all ML models with the trained models using synthetic data.

The best performing model among all experimental settings is XGBoost that used the selected features based on forward features selection techniques, numerically the performance 20%, 7% and 5% higher than the best performing model trained on original data (Table 13) in terms of Kappa, Accuracy and AUC, respectively. Based on AUC score, XGBoost model on real data with backward feature selection technique achieved highest performance (numerically, 0.9438 (Table 13)) and on contrary the XGBoost model trained on synthetic data generated by CTGAN is way more higher than the the performance (numerically, 0.9885) on same model trained with original data.

The superiority in modeling thermal comfort preference with synthetically generated data is visualized in Fig. 31 and 32 using bar and line chart. In both figures, we highlight the difference in achieving the higher performance in terms of most significant evaluation metrics AUC between model trained with original data and with synthetic data for all 24 experimental settings. In Fig. 31, the black colored bar denotes the performance for trained models on synthetically generated data by CTGAN and the gray colored bars represent the performance for models on original data. Similarly, the orange and blue colored line in Fig. 32 represent the similar performance of models with and without synthetic data.

From both the figures we can see that, the models trained with synthetic data demonstrated higher performance compared to all experimental settings except two models, decision tree for Ch2-based and correlation-based feature selection technique. Fig. 32 clearly illustrates the performance improvements after integrating CTGAN based synthetic data generation techniques. However, based on the facts and findings discussed above we can conclude that feature selection on synthetic data can achieve higher performance in predicting thermal comfort preference and that can be used to calibrate the indoor environment. Hence,

Table 15: Performance in modeling PTC preference compared to with baseline.

Sub. ID	Model	[CTGAN +FS]			Liu et al. [191]		
		Kappa	Accuracy	AUC	Kappa	Accuracy	AUC
1	XGB	**0.2791**	0.5294	**0.8315**	0.17	**0.56**	0.68
	RF	0.2718	0.5490	0.7994			
2	XGB	0.3049	0.6666	**0.8202**	**0.51**	**0.74**	0.75
	RF	0.1898	0.5833	0.8065			
3	XGB	**0.7816**	**0.8785**	**0.9823**	0.50	0.77	0.86
	RF	0.7521	0.8598	0.9778			
4	XGB	0.1684	0.5393	**0.8059**	0.07	**0.86**	0.73
	RF	**0.2014**	0.5505	0.7904			
5	XGB	**0.5901**	**0.7528**	0.9311	0.30	0.69	0.79
	RF	0.5191	0.6853	**0.9327**			
6	XGB	**0.3447**	**0.5500**	**0.7973**	0.17	0.53	0.63
	RF	0.2595	0.4875	0.7970			
7	XGB	**0.5399**	**0.7230**	**0.9289**	0.37	0.69	0.77
	RF	0.5334	0.7153	0.9245			
8	XGB	**0.6666**	**0.8974**	**0.9816**	0.33	0.88	0.84
	RF	0.3960	0.7350	0.9610			
9	XGB	**0.3768**	0.6547	**0.9153**	0.21	**0.79**	0.81
	RF	0.3741	0.7261	0.9020			
10	XGB	**0.5430**	**0.7294**	**0.8947**	0.18	0.63	0.67
	RF	0.4430	0.6588	0.8555			
11	XGB	0.43557	0.6590	**0.9149**	0.37	**0.79**	0.79
	RF	**0.4590**	0.7045	0.9109			
12	XGB	**0.4197**	**0.6545**	**0.8756**	0.18	0.63	0.76
	RF	0.3243	0.6363	0.8408			
13	XGB	**0.9000**	**0.9545**	0.9477	0.41	0.75	0.83
	RF	0.7928	0.9090	**0.9643**			
14	XGB	0.1793	0.4018	**0.9416**	0.04	**0.80**	0.75
	RF	**0.1917**	**0.4392**	0.9030			

this might provide more occupant-friendly environment that can be both healthy and energy efficient in smart home.

7.4.7 Performance on PTC Preference Prediction

To predict occupants' PTC preferences, we trained two ML models, including XGBoost and Random Forest, which are the two best-performing models on real and synthetically generated data. We considered the selected features by applying backward feature elimination techniques and trained both models for each subject separately on synthetically gener-

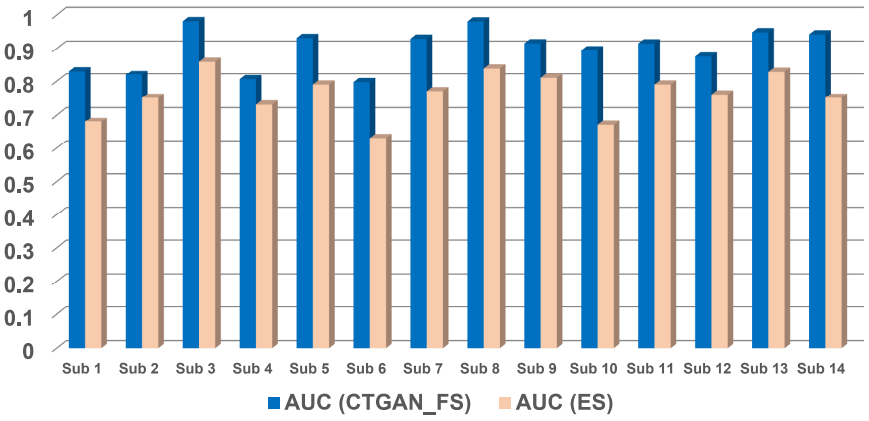

Figure 33: Performance comparison with existing study in terms of AUC

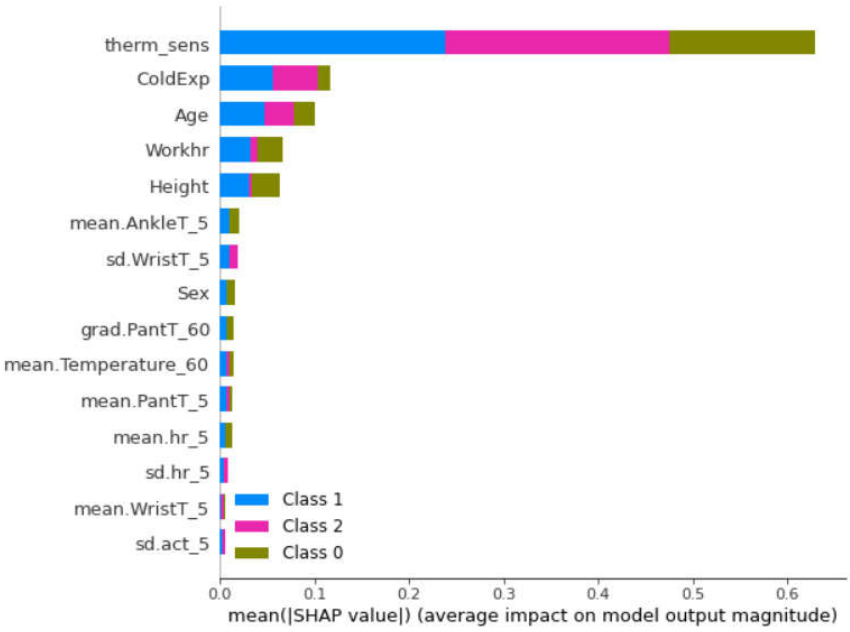

Figure 34: Interpretation of PTC model with feature selection using SHAP values

ated data leveraging CTGAN. The experimental results in predicting PTC preferences for all 14 different subjects are presented in Table 15. We also compared the results with existing work [191], where they trained classical ML models on the same dataset. The best PTC preference prediction performance among multiple ML models for every subject [191]

is reported in the right-hand side of Table 15.

Our trained models with the selected features on synthetic data by CTGAN outperformed the models in related work in terms of Cohen's Kappa, except for one subject (Subject 2). In terms of the most important metric, AUC, our method, combining synthetic data generation and feature selection techniques, significantly outperformed existing works for all 14 different subjects. AUC is the metric that can better measure a classifier's performance for imbalanced data distribution.

Liu et al. [191] applied multiple classical ML models on the annotated data, considering all features. As we mentioned and analyzed earlier, some features are correlated, redundant, and irrelevant. We observed that our feature selection techniques identify the best sets of features related to modeling PTC preference. Compared to prior research, their model might suffer from irrelevant features that might mislead the classifiers.

One of the major problems in training the PTC preference model is the inadequacy of sufficient data samples. We tackled this problem by introducing CTGAN to generate high-quality synthetic data, and utilizing those data, we trained the models efficiently. Hence, the model has more data points about particular subjects and can learn better to predict thermal comfort preference more accurately. The dataset was imbalanced, CTGAN also solved that problem and mitigated the possible bias-related issues in predicting thermal comfort. Based on the evaluation metrics Kappa and AUC, both recommended due to the data imbalance issue, our introduced models demonstrated significant improvements. Therefore, we can say that applying CTGAN for data generation for personal thermal comfort with feature selection is an effective approach that can be an effective combination to achieve high performance in PTC prediction.

To point out the performance differences compared to previous work,

we present a comparison with related work as a bar chart in Fig. 33 in terms of performance based on AUC. AUC (CTGAN+FS) and AUC (ES) denote the performance of our proposed method combining CTGAN and feature selection (FS) and existing study (ES), respectively, in terms of AUC. We can also observe from the figure that our method outperforms prior work for all 14 subjects' thermal comfort preference prediction. To the best of our knowledge, this research is one of the few studies that conduct an extensive study on the impact of synthetic data generation by CTGAN and effective feature selection techniques in PTC preference prediction.

7.4.8 Model Interpretability

To understand the priorities in decision-making of PTC model, we applied one of the successful method, shapely additive explanation (SHAP) [195]. The feature interpretation using SHAP is presented in Fig. 34. Note that, we conducted this interpretability experiments with the selected features by applying correlation-based feature selection technique. The figure shows the 15 most important features that contributed most to make the decision of our proposed PTC model.

The top two most important features that the model takes into account are thermal sensation and cold extremity experience. It makes sense that thermal comfort preference should be dependent on these two features most. However, the next three features are age, total working hours per day and height of the subjects. Next, the skin temperature at ankle and wrist contributed the PTC model in decision making. Interestingly, subjects' sex is also an important feature that is considerable in predicting the PTC. In turn, body proximity temperature, outdoor temperature, and heart rate is also considerably important features in modeling PTC preference. This is in line with the feature selection results that we presented in Table 12.

7.5 Conclusion and Future Work

This paper proposes a thermal comfort preference prediction method that combines a two-step process involving synthetic data generation using CTGAN and the selection of the best set of features by filtering out irrelevant and noisy features using multiple feature selection approaches. The results on a wide range of experimental settings demonstrated state-of-the-art performance and significantly outperformed existing known related work.

We observed that the ML models trained on synthetic data generated by CTGAN can predict better PTC preference than on original data samples. In addition, the introduction of a series of feature selection techniques helps filter out irrelevant features in modeling PTC preference prediction tasks. The interpretability of the model with SHAP demonstrated that the important features also overlap with the selected features. Since the PTC preference prediction task needs a substantial amount of data samples per subject to train the model efficiently, the incorporation of an effective data generation technique can save both data collection costs and associated time. The findings with feature selection indicate not to collect unnecessary data from the subject and environment and hence it might also save potential cost in sensor-based data collection cost.

In the future, we plan to introduce explainable artificial intelligence (XAI) on a large scale to provide a human-centered explanation so that occupants can understand the reason behind specific indoor parameter changes related to thermal comfort in smart homes. Since the PTC preference prediction model will be used in the smart home to control the indoor environment, an exciting extension of this work would be to conduct a user study to design a new collaborative interface for general users in human-computer interaction (HCI) perspective so that they can also be included in the loop of the smart heating system.

Acknowledgment

This research has been funded by the EU Horizon 2020 Marie Sklodowska-Curie International Training Network GECKO, Grant number 955422.

Part III

Explainable models in Business

8 Introduction

Our research delves into the nuances of explainability techniques across diverse application scenarios. To shed light on this, we have selected a unique e-commerce application-product backorder prediction. This choice is significant as it introduces a new explainable deep learning model in the business area, specifically tailored to the needs of inventory management stakeholders. We have also made efforts to present the explanations in various forms and visualizations, enhancing the stakeholders' understanding of their product.

This area somewhat differs from the previous applications presented in the last Part II. The users of both systems are different, and their expertise in understanding and comprehending explanations varies. The users of smart home applications are general people, and the stakeholders for product backorder prediction systems are experts in inventory management. So, the degree of comprehension of the explanation is different. In one of the challenges, we discussed how the variability of user expertise affects the understanding of the explanation.

With a focus on our objectives, we propose a new explainable convolutional neural network-based product backorder system. Our detailed experiments on a standard dataset have shown a significant improvement in accurate prediction compared to the known-related works. This reassures the stakeholders of the robustness and reliability of our new model, even in the face of a biased dataset with extreme imbalance, as evidenced by our efficient prediction of backordered products and higher AUC.

The explanations we have generated, by incorporating SHAP and training a surrogate model with LIME, are not just comprehensive but also highly effective. They are presented in various forms and visualization techniques, empowering the stakeholders to not only understand but

also take corrective action to maximize their revenue potential.

9 Explainable Product Backorder Prediction Exploiting CNN: Introducing Explainable Models in Businesses

The content of this chapter has been published as a full paper in *Electronic Markets* by Springer Nature. The information of the published paper is given below:

Information of Article: Md Shajalal, Alexander Boden and Gunnar Stevens. 2022. Explainable Product Backorder Prediction Exploiting CNN: Introducing Explainable Models in Businesses. *Electronic Markets*, Springer-Nature 32, 2107–2122. https://doi.org/10.1007/s12525-022-00599-z (*Reproduced with permission from Springer Nature*)

Abstract

Due to expected positive impacts on business, the application of artificial intelligence has been widely increased. The decision-making procedures of those models are often complex and not easily understandable to the company's stakeholders, i.e. the people having to follow up on recommendations or try to understand automated decisions of a system. This opaqueness and black-box nature might hinder adoption, as users struggle to make sense and trust the predictions of AI models. Recent research on eXplainable Artificial Intelligence (XAI) focused mainly on explaining the models to AI experts with the purpose of debugging and improving the performance of the models. In this article, we explore how such systems could be made explainable to the stakeholders. For doing so, we propose a new convolutional neural network (CNN)-based explainable predictive model for product backorder prediction in inventory management. Backorders are orders that customers place for products that are currently not in stock. The company now takes the risk to produce or acquire the backordered products while in the meantime, customers can cancel their orders if that takes too long, leaving the company with unsold items in their inventory. Hence, for their strategic inventory management, companies need to make decisions based on assumptions. Our argument is that these tasks can be improved by offering explanations for AI recommendations. Hence, our research investigates how such explanations could be provided, employing Shapley additive explanations to explain the overall models' priority in decision-making. Besides that, we introduce locally interpretable surrogate models that can explain any individual prediction of a model. The experimental results demonstrate effectiveness in predicting backorders in terms of standard evaluation metrics and outperform known related works with AUC 0.9489. Our approach demonstrates how current limitations of predictive technologies can be addressed in the business domain.

Keywords

eXplainable Artificial Intelligence (XAI), Backorder Prediction, CNN, Local Explanation, Global Explanation

9.1 Introduction

Due to their superior predictive performance, complex machine learning and deep neural network-based models have received high attention and are widely exploited in the business domain [35, 137, 69] along with other fields including image processing [143], health [228, 33] and bioinformatics [52, 186]. The tasks of those technologies range across different application areas including supply chain management, credit risk prediction [212, 51], detection of fraud credit card transaction [53, 242] and marketing campaigns in retail banking [171].

Generally, artificial intelligence (AI) techniques employ a huge size of training data for making predictions. While there is a huge interest in such predictions in various business domains [244], one of the major problems of complex machine learning models is that they are very difficult to understand [3, 4, 306]. Several Methods using induced ordered weighted averaging (IOWA) adaptive neuro-fuzzy inference system (ANFIS) can deal with multidimensional data to predict the quality of service and hence it help stakeholders in the decision-making process [131, 129, 130]. As decisions often depend on a huge number of model parameters [16], machine learning and deep learning techniques are like black-boxes or magic boxes to the general users (and often even for developers). The higher the accuracy of a complex machine learning model, the more opaque the models tend to become [244]. This opaqueness leads to a situation where users might question the predictions, because they are unable to understand the underlying decision making processes (i.e. the reasons for why a maybe counter-intuitive recommen-

dation has been given) [23].

User acceptance is generally one of the main barriers for the success of technologies in companies. As AI-based recommendations can potentially have a huge impact on operational as well as strategic decisions in companies, it seems to be beneficial if users or consumers of AI models could better understand why those recommendations have been made [202]. Apart from increasing trust in AI-recommendations, having factual explanations of a certain decision would also help users to learn about the field of application (for instance gaining a better understanding in the importance or non-importance of certain factors for business decisions) [103]. In addition, according to the general data protection regulation (GDPR) by the European Union, EU citizens have the right to receive explanations about AI-based decisions, for instance if an AI recommendation affects credit worthiness or insurance rates [202, 83].

In this research, we propose a novel explainable predictive model for product backorder prediction. A backorder is a situation where customers can order a product even though that particular product is out of stock at the time when the order is placed [119, 226]. Basically, its an order to a future inventory, going along with contingencies as time of delivery can vary and is not definitely known. Backorders are especially common for items that are highly popular. While for some items such as the latest flagship Apple iPhone, such events are quite common, it can be very unpredictable for other types of products. When retail companies order high amounts of products based on backorders, they risk their reputation if they are unable to keep the expected delivery dates. Another risk is that customers can cancel their orders because they don't want to wait any longer or found another retailer where the product is in stock, leaving the company with excess products in their inventory. Here, predictive models can help to tackle these challenge by predicting the probability whether a certain product will be

backordered or not, giving companies more time to plan and supporting them in their inventory management. In related works, researchers have proposed complex machine learning based methods to predict future product backorders. The predictive models include the application of support vector machine, XBoost, ensemble classifier and deep neural networks [134, 119, 185, 286].

However, the mere prediction of future backorders only solves part of the problem. Suppose you are responsible for a particular inventory management system at a retail company. When you are notified that the AI model decided that a particular product is going to be backordered in the near future, what will you do? *Would you increase the inventory level (i.e. obtaining more products in advance)? Would you change any policy (negotiating with suppliers about faster transit times, lead times etc)?* If you increase the inventory level, *how many products would you order, assuming that some would surely be cancelled?* For taking these decisions, you would need to understand the reasons for the prediction. Hence, our approach tries to provide insights into the factors that contribute to a certain prediction, helping users to adapt their strategies accordingly. Our paper contributes in the following ways:

- We proposed a new CNN-based model for backorder prediction. Since backorders are rare events in inventory management systems, it is a challenging task to identify them. Their rarity leads to an extremely imbalanced distribution within datasets. Often, the percentage of the backodered samples is less than 0.01% (specifically 0.007%, [79]). To address this data imbalance, we incorporated an adaptive synthetic oversampling (ADASYN) technique that generates synthetic samples for a minority class. The results, based on diverse experimental settings and comparison with existing known related works, illustrated that our method achieved better prediction performance achieving a new state-of-the-art meth-

ods performance in terms of standard evaluation metrics.

- To provide an overall insight of the predictive model's decision-making priorities, we investigate the impact of different attributes of an order in the predictive models. We introduce an XAI technique, namely SHAP (Shapely additive explanations), that can interpret and/or explain the predictive model to identify the most important attributes of the decision making. Hence the stakeholders are enabled to better understand the model's decision-making priorities and consider that when they have to work with such technologies.
- By explaining specific predictions, our method can answer why a particular product will be backordered or not. Every order has different feature's values which are considered to make predictions. Therefore, we trained a local interpretable surrogate model employing LIME (local interpretable model-agnostic explanations), and present explanations for an individual prediction to answer the question "*why has this specific decision been made?*" Hence, stakeholders can not only assess the models' priorities in general, but also analyse singular decisions to better understand them.

The organization of the rest of this paper is as follows: Section 9.2 summarises related works on predicting product backorders. We present a brief discussion about different XAI terminologies in section 9.3. In Section 9.4, we present our method for predicting future product backorders and the explanation generation techniques. The predictive performance of our proposed CNN-based method and performance comparison with classical machine learning classifiers and known related works are presented in detail in section 9.5. The decisions of complex machine learning and deep learning models are explained through different types of explanations both for models' priorities as well as specific predictions in section 9.6. Finally, section 9.7 concludes our proposed methods and

findings of this study by discussing the prospects of introducing XAI technology in the business domain.

9.2 Related Work

This section presents the discussion of related research on backorder prediction and explainable artificial intelligence in supply chain management. Existing works proposed different models to predict plausible future backorders in inventory management systems. Based on the types of techniques used, the predictive models can broadly be classified into two categories: i) Classical machine learning classifiers and ii) Deep learning-based predictive models. In the former category, the classifiers include support vector machine [119], gradient boosting [226, 79], decision trees, and random forests [134]. The deep learning-based models employed recurrent neural networks (RNN) [185], deep autoencoders [262], as well as deep neural networks (DNN) [286].

Islam et al. [134] proposed a method to predict future backorders by applying distributed random forest and gradient boosting classifiers. They introduced a ranged-based approach to cope with the numerous types of real-time data. However, they did not include some features of the samples such as features related to inventory level, previous sales, future sale forecasting, and lead time. A profit-maximizing function based approach is introduced by Hajek and Abedin [119].

Table 16: The summary of existing study on product backorder prediction

Research Paper	**Contributions**	**Explainabaility**
Islam & Amin [134]	• Applied Distributed Random Forest (DRF) and Gradient Boosting Machine • Employed SMOTE oversampling	• Inherent (Global feature importance) • No local explanation to understand particular decision
Hajek & Abedin [119]	• A genetic algorithm-based profit maximizing prediction system • Applied classical ML classifiers	• Not Explainable

Shajalal et al. [286]	• Proposed a deep neural network-based prediction model • Combined random undersampling and synthetic oversampling to overcome class imbalance problem	• Not Explainable
Li et al. [185]	• Applied recurrent neural network (RNN) based predictive model • Exploited SMOTE, ADASYN, and random undersampling for balancing the dataset	• Not Explainable

Saraogi et al. [262]	• Proposed deep autoencoder based model for backorder prediction • Used unsupervised approach rather than supervised one	• Not Explainable
Ntakolia et al. [226, 225]	• Introduced classical machine learning model to predict backorder • Incorporated interpretability to understand decision making	• Interpreted the global feature importance to explain the model • No local explanation for particular decisions

Santis et al. [79]	• Exploited classical machine learning classifier including gradient boosting and ensemble model • Used bagging to overcome data imbalance problem	• Not Explainable
Lawal & Akintola [174]	• Applied recurrent neural network (RNN) based predictive model • Exploited SMOTE, ADASYN, and random undersampling for balancing the dataset	• Not Explainable

They aligned their profit maximization function with classical machine learning classifiers. The performance of their methods demonstrated how much profit can be increased by predicting future backorders. An explainable classical machine learning-based method is proposed by [226]. Their method applied several classifiers such as random forest, XGBoost, SVM, etc. They also applied shapely additive values to present the global explanations to interpret the models. Similarly, Santis et al. [79] also used different classical classifiers. The performance

of deep learning approaches is comparatively better than the classical classifiers. Shajalal et al. [286] proposed a deep neural network (DNN) based backorder prediction model. Inspired by the success of deep learning classifiers, Li et al. [185], Saraogi et al. [262] and Lawal and Akintola [174] applied deep auto-encoder, a recurrent neural network-based classification models.

Backorders are not a common scenario in inventory management systems. In turn, the number of non-backordered items is much larger than the backordered ones. Hence, real-time data collected from any inventory system will be strongly imbalanced, leading to challenges in predicting future backorders on that basis. In this particular task [185], the ratio between majority (non-backordered) and minority (backordered) samples is 100:0.007. In the case of an imbalanced dataset, the classifiers might learn the pattern with potential bias. That is why different under-sampling, oversampling, and class weight-based approaches are common to balance the dataset and bias [119]. Randomly duplicating the minority samples or randomly discarding the majority samples has also been applied to balance the dataset [62]. But randomly duplicating the minority samples will increase redundant samples and hence the model might be biased. Therefore, generating synthetic minority samples based on the Euclidean distance is a popular approach to balance the dataset. This method is called SMOTE (Synthetic Minority Over-sampling Technique) [62]. The combination of SMOTE and random under-sampling has been applied by Hajel & Abedin [119] and Shajalal et al. [286, 273]. Li et al. [185] applied different balancing techniques including SMOTE, ADASYN (Adaptive Synthetic Sampling) [123] and random under-sampling. Bagging [44] is also applied for the same purpose by Santis et al. [79]. Table 16 summaries the existing methods for predicting product backorders. However, to the best of our knowledge, none of the studies applied XAI to interpret their machine learning

Table 17: Existing research gaps in explainable product backorder prediction and our steps to fulfil the research gaps.

Issue	**Research Gaps in Literature**	**Our Contributions**
Performance	Most existing methods suffer from low performance in modelling product backorder prediction due to extreme data imbalance problem. The majority of prior studies applied classical ML methods.	We proposed a novel CNN-based prediction model with the ADASYN technique that achieved new state-of-the-art performance.
Model's Interpretability	Lack of interest in applying XAI to explain the predictive model's decision-making priorities. Hence the existing models can be seen by the stakeholders as black-box.	We introduced shapely additive explanation (SHAP), one of the most successful XAI techniques to explain the models' global priorities that help stakeholders in sense-making about the working strategy of the predictive models.
Local Explaiability	No existing works explain the specific prediction to answer "*why has this specific decision been made?*" (i.e., why a certain product is going to be backordered?)	We exploit LIME and SHAP to explain specific predictions about why a particular product is assumed to be on backordered or not. These techniques can explain which features/attributes are responsible for a particular decision. Hence the stakeholders are enabled to take steps to overcome future backorder and reduce the company's loss.

model except [226].

In our paper, we propose a convolutional neural network framework-based model that outperformed different classifiers including classical

and deep learning-based models in backorder prediction. Ntakolia et al. [226] interpreted only classical models mainly with global explanations. Our method integrated explainable artificial intelligence that generates global explanations for the classical and deep learning-based prediction model. Though the global interpretation is useful to illustrate the general mechanisms and behavior of the model, it can not explain a particular prediction. We introduced a model applying shapely additive explanation [195] and local interpretable model-agnostic explanation [244] to interpret the overall model and local specific decisions.

To clearly illustrate the research gap in the existing literature and our research focus, we present a comparative analysis in Table 17.

9.3 XAI Terminology

In this paper, we employed two XAI techniques, namely shapely additive explanations (SHAP) and local interpretable model-agnostic explanations (LIME). Here, we present the background and working principle of these two techniques.

9.3.1 SHapely Additive exPlanation

Lundberg and Lee [195] first proposed a unified approach to explain and interpret the prediction of machine learning models. The explanations basically illustrate the contributions (positive and negative importance or influence) of different features for the predicted decision of a particular sample x. The overall feature importance of different features of the whole model can also be interpreted as global explanations. In that case, the importance score resembles the weight of features as in the linear model. The SHAP values represent the importance of the features. The explanation of every single prediction can be seen as a vector of shap values. The same representation is used to interpret the overall model.

For a given instance x, the explanation using SHAP can be defined as

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_j z'_j, \tag{9.1}$$

where g is denoted as the explanation model. The vector for simplified features, known as the coalition vector is represented by z' ($z' \in \{0,1\}^M$). The 1 represents that features' values are the same as the original instances and vice-versa. The attribution of particular features j of the instance x is denoted by ϕ_j which is a real number. The higher the value of ϕ_j, the more important the feature j. The ϕ_j is computed based on Shapely values [224], a game-theoretic approach that identifies and detects the contribution of all players in a collaborative game. The collaborative game with multiple players is analogue to the prediction of the instance having multiple features. In turn, applying this game-theoretic approach we can examine the contribution of each feature to a particular decision. For a given feature vector x' and a predictive model f, the computation is done as follows:

$$\phi_i(f,x') = \sum_{z' \subseteq \{x'_1, x'_2, \ldots, x'_n\} \setminus \{x'_i\}} \frac{(|z'|)!\ (M-|z'|-1)!}{M!} \cdot \left[f(z' \cup x'_i) - f(z') \right] \tag{9.2}$$

The subset of the features employed by the model is denoted as z'. x' is the vector with features values to be explained and can be defined as $[f(z' \cup x'_i) - f(z')]$ and M is the number of features. The prediction by the model f is denoted by $f(z')$. Moreover, SHAP values are computed by a standard game-theoretical approach and utilized Shapely values to have a unified interpretable model with fast computation. More mathematical and technical details for SHAP can be found in the study published by Lundberg et al. [195] as well as in [224].

9.3.2 Local Interpretable Model-agnostic Explanation

LIME mainly provides model-agnostic explanations based on local surrogate models. Ribeiro et al. [244] first introduced this approach for training a local surrogate model instead of a global model for providing explanations for a particular prediction. LIME employed a new local dataset containing the permuted samples with corresponding predictions to train the local interpretable surrogate model. This surrogate model is then used to explain individual predictions. The model is considered as a approximation of the original complex, black-box predictive model. The computation of the surrogate model can be defined as follows:

$$\xi(x) = \underset{g \in \mathbf{G}}{\arg\min}\, L(f, g, \pi_{x'}) + \Omega(g) \tag{9.3}$$

The explanation model for a particular instance x and the explanation family are represented by g and G, respectively. The original model is denoted by f and L is the loss function. The complexity of the model can be defined by $\Omega(g)$. LIME is useful to explain specific decision predicted by the model (i.e., local prediction).

9.4 Explainable Product Backorder Prediction

The overview of our proposed explainable product backorder prediction framework is depicted in Fig. 35. We first apply preprocessing step to handle the missing values, converting qualitative variables into quantitative ones and normalizing the values in a similar range. Next, we apply our proposed convolutional neural network-based backorder prediction model to classify the product. Finally, we introduce explainable AI techniques to explain both global, model agnostic aspects as well as individual decisions with the intent to make the inventory manager understand better why his or her backorder prediction system acts as it does.

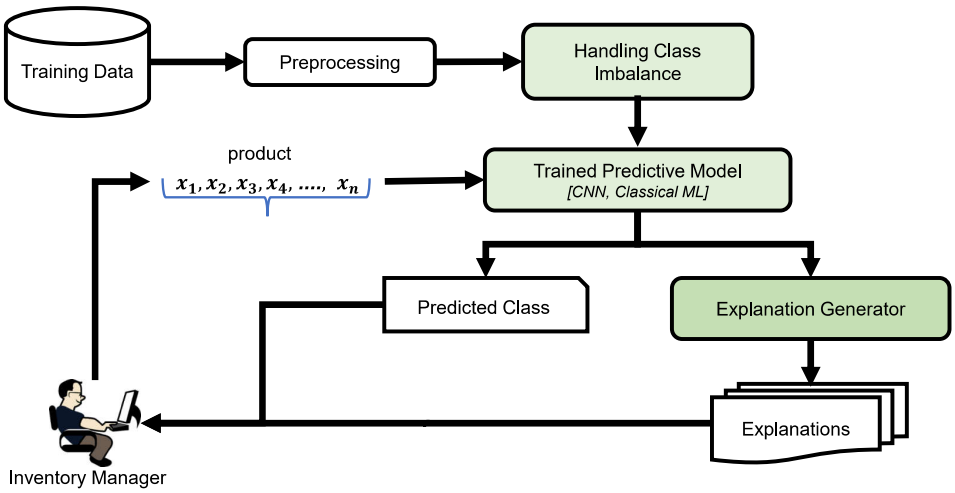

Figure 35: Proposed explainable backorder prediction approach

9.4.1 Preprocessing and feature analysis

In our dataset, each particular sample has 21 different features/attributes including current inventory, lead time, forecasting for a different time, sales performance, different risk flags. The details of the dataset are presented in section 9.5.1. The value of different features is varied widely among binary, quantitative, qualitative, and categorical. In this step, all the feature values are transformed into a real number. The missing values are handled by filling them in with the median of other samples' values. A normalization technique is then applied to convert each feature value into a certain range [0,1]. Here, we applied the most widely recognized MinMax normalization technique.

However, a dataset having highly correlated features is not suitable for applying classification methods. We investigate to see whether any high correlated features are available, exploiting the Pearson correlation coefficient measure for this purpose. According to the findings, we observe that there are no features with a high correlation ($\rho > .80$). Hence, the dataset should now be suitable for our purpose of backorder predictions.

Algorithm 1 ADASYN: Adaptive Synthetic Oversampling

procedure ADASYN_OVERSAMPLING(D_{train})
  $d = \frac{m_{min}}{m_{maj}}$ ▷ Finding the imbalance ratio
  if $d < d_{th}$ **then**
    $G = (m_{maj} - m_{min}) \times \beta$ ▷ Number of synthetic examples need to generate
    for each $x_i \in X_{min}$ **do**
      N_{kn} = K_Nearest_Neighbors(x_i)
      $r_i = \Delta_i / K, \quad i = 1, ..., m_{min}$
    end for
    Normalize r_i such that $\hat{r}_i = r_i / \sum_{i=1}^{m_{min}} r_i$
    $g_i = \hat{r}_i \times G$ ▷ Number of synthetic example per sample x_i
    for 1 to **do** g_i
      $x_{zi} = Random\_choice(X_{min})$
      $s_i = x_i + (x_{xi} - x_i) \times \lambda$
    end for
  end if
end procedure

9.4.2 Handling class imbalance with ADASYN

As we noted earlier, a product backorder scenario is a rare event that leads the dataset to be extremely imbalanced. Therefore, we employed one of the efficient synthetic oversampling methods, ADASYN (Adaptive Synthetic oversampling) [123] to balance the dataset. Considering the difficulty level of learning, ADASYN generates synthetic minority class examples utilizing the weighted distribution. ADASYN focused on generating more synthetic minority class examples for those minority samples that are harder to classify. Given a training dataset, D_{train} with N number of samples where each sample is denoted as x, y, the vector x is represented by a K dimensional vector containing different attributes of an ordered product and y is the binary value that indicates the label (0 for non-backordered and 1 for backordered one).

Let m_{min} and m_{maj} be the number of examples of minority class and majority class, respectively such that $m_{min} + m_{maj} = N$, and in this backorder prediction task $m_{min} << m_{maj}$. ADASYN oversampling techniques generate

synthetic minority class examples to balance the dataset according the algorithm illustrated in Algo. 1.

It first calculates the degree of imbalance d and then, depending on the tolerated imbalance ratio, computes the number G that denotes the number of synthetic minority class examples needed to be generated. Here $\beta \in [0,1]$ indicates the desired bleaching ratio, $\beta = 1$ indicated that the dataset will be fully balanced. For each minority example x_i, ADASYN then calculate the ratio r_i applying K-nearest neighbors with Euclidean distance, where Δ_i is the number of nearest neighbors of x_i. Using the normalized ratio $\hat{r}_i$, then it computes the number of synthetic examples for each minority examples x_i. Finally it generates the synthetic minority class examples applying the distance vector and the random number λ.

9.4.3 Convolutional Neural Network-based Prediction Model

Inspired by the success of the convolutional neural network (CNN)-based models in computer vision, natural language processing and other classification tasks, we proposed a 1-dimensional CNN classifier to predict product backorder in advance. The structure of our proposed CNN-based predictive model is illustrated in Fig. 36.

Our CNN-based predictive model has two convolutional hidden layers with batch normalization, max-pooling, and dropout layers. To extract unique and low-level features, the max-pooling layers are exploited. In addition, max-pooling makes the computation faster by reducing the dimension and parameters [323]. Moreover, it reduces the variance. Then we utilized one flattened layer followed by three dense layers with dropout layers. To overcome the over-fitting problem, dropout layers are applied to randomly drop some neurons in the training process for regularization [165, 301]. The parameters and activation functions in different layers of convolutional neural networks are summarized in Table 18.

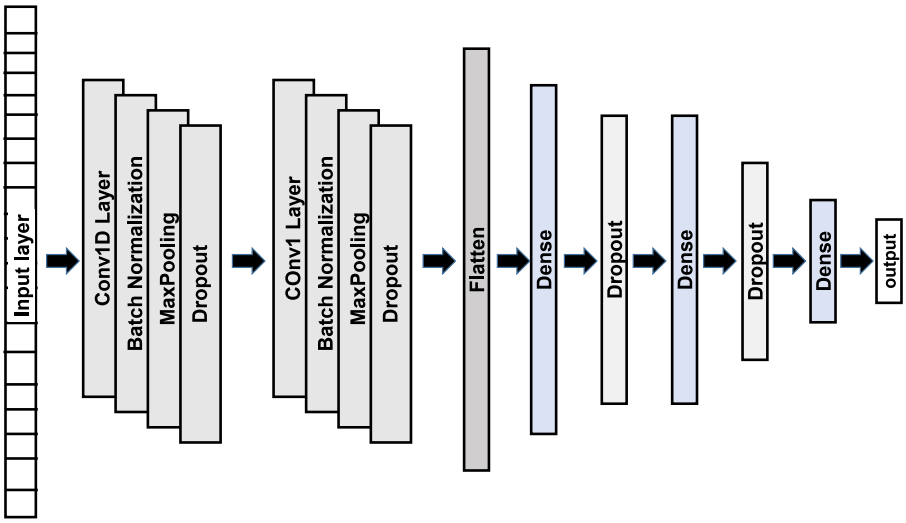

Figure 36: Structure of our proposed convolutional neural network-based backorder prediction model

In the convolutional layers and all hidden dense layers, we employed the *Relu* [240] activation function. Finally, *Sigmoid* [240] activation function is applied in the output layer.

Table 18: The summary of different layers with parameters and activation functions.

SL	Layer	Input/Output	Activation
1.	Conv1D	(20,32)	*Relu*
2.	Batch Normalization	(20,32)	-
3.	Max Pooling	(10,32), stride=2	-
4.	Dropout	(10,32)	-
5.	Cov1D	(9,64)	*Relu*
6.	Batch Normalization	(9,64)	-
7.	Max Pooling	(4,64), strid=2	-
8.	Dropout	(4,64)	-
9.	Flatten	-	-
10.	Dense	64	*Relu*
11.	Dropout	64	-
12.	Dense	32	*Relu*
13.	Dropout	32	-
14.	Dense	1	*Sigmoid*

9.5 Experiments and Evaluation

9.5.1 Dataset collection and evaluation metrics

This section presents the details of dataset that is leveraged to conduct experiments using our proposed method. We also present a brief discussion about the evaluation metrics considered to measure and validate the performance.

Dataset: We carried out a wide range of experiments to validate the performance of our methods on a publicly available benchmark dataset called "*Can you Predict Product Backorder*[7]". The dataset has an 8 weeks inventory of historical data. The brief statistical summary of the dataset is depicted in Table 19. The numerical figures in Table 19 illustrate that

Table 19: Brief statistical summary of the dataset

No of Samples	No of Positive Samples	No Nagative Samples	Imbalance Ratio
1,929,936	13,981	1,915,954	1:137

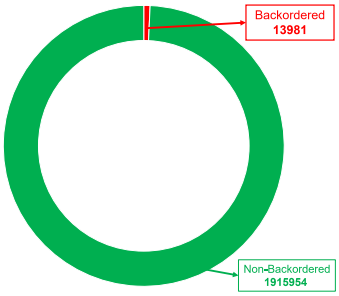

Figure 37: Distribution of backordered and non-backordered samples

the number of backordered (positive) samples is much lower than the number of non-backordered (negative) samples. Hence, the ratio (1:137) indicates that this dataset is an extremely imbalanced one. For a better understanding of why this is a challenging problem, we illustrated the distribution of backordered (positive) and non-backordered (nega-

[7] https://github.com/rodrigosantis1/backorder_prediction

tive) samples using a doughnut chart in Fig. 37. There are 22 features for each sample and the attributes/features include current inventory, transit time, quantity, forecasting, and different risk flag. The list of features with a brief description is depicted in Table 20.

Table 20: Description of different features/attributes of a particular order

Feature f_i	**Explanation**
Current Inventory f_1	Current inventory level of component
lead Time f_2	Registered transit time
Transit Quantity f_3	Product amount in transit
Sale Forecasts f_4, f_5, f_6	Forecasting stock amount for upcoming 1, 3, 6, 9 months
Sales f_7, f_8, f_9, f_{10}	Amount of sold product in last 1, 3, 6, 9 months
Reco. Stock Amount f_{11}	Recommended amount (amount) in stock
Overdue f_{12}	Overdue parts from source
Performance f_{13}, f_{14}	Performance of the product in last 6 months and 12 months
Stock Overdue f_{15}	Overdue Stock amount for orders
Risks $f_{16}, f_{17}, f_{18}, f_{19}, f_{20}, f_{21}$	Different general risk flags
Label Y	Product went on backorder or not

Evaluation metrics: Generally, the performance of any classification method is measured based on the common evaluation metrics including *accuracy*, *precision*, *recall* and f_1*-score*. The confusion metrix is used to compute those metrics. However, the backorder prediction dataset is extremely class imbalanced, and the above mentioned evaluation metrics are not enough to validate the performance of any classifier on a imbalanced dataset. Therefore, we employed *accuracy*, *AUC* (**A**rea **U**nder the **C**urve) and *ROC* (**R**eceiver **O**perating **C**haracteristics) curves to measure and visualize the performance of our proposed backorder prediction method. The accuracy score is calculated by using the measures from

confusion metrics as follows:

$$Acc = \frac{tp + tn}{tp + fp + fn + tn}, \tag{9.4}$$

where *tp*, *fp*, *fn* and *tn* denote the number of classified samples as true positive, false positive, false negative and true negative, respectively.

AUC is one of the most efficient metrics to measure the performance of any classification model on imbalanced data. The *AUC* is calculated as follows:

$$AUC = \frac{1 + P - F}{2}, \tag{9.5}$$

where P is the precision of the classifier and F is the false positive rate. The details of the these metrics can be found elsewhere in the published study by Chawla et al. [62] and Santis et al. [79].

9.5.2 Prediction Performance

We conducted a wide range of experiments with multiple settings to illustrate the performance of our backorder prediction methods. Since our major goal in this study is to introduce the explainability of in backorder prediction, we first applied classical machine learning and a deep neural network-based classification approach. Then, we exploit XAI technologies (SHAP and LIME) to explain the model's priorities and individual prediction. Classifiers from classical machine learning including decision tree, support vector machine, gradient boosting, etc., were applied. All experimental settings can be broadly classified into three different types based on the chosen dataset balancing strategy, classical ML, and deep learning. In all experimental setups, we applied two different dataset balancing techniques ADASYIN and SMOTE [62]. Based on the predictive models, we report the experimental results in two categories, classical and deep classifiers.

Table 21: Performance of classical machine learning models in terms of *accuracy* and *AUC*. The best result is in **bold**.

Predictive Model	**Balancing Tech.**	**Accuracy**	**AUC**
Decision Tree	ADASYN	0.9366	0.8134
	SMOTE	0.9338	0.8111
Support Vector Machine	ADASYN	0.8511	**0.8711**
	SMOTE	0.8455	0.8696
Gradient Boosting	ADASYN	**0.9548**	0.8298
	SMOTE	0.9538	0.8288
Gaussian Naive Bayes	ADASYN	0.7947	0.8153
	SMOTE	0.7836	0.8142
K-nearest Neighbor	ADASYN	0.8970	0.8498
	SMOTE	0.8977	0.8507

The prediction performance of all experimental setups using classical machine learning is presented in Table 21. The results conclude that the ASASYN balancing technique is more efficient and achieved higher accuracy as well as AUC than SMOTE in most of the experimental setups. In turn, we can conclude that for the backorder prediction task, our introduced ADASYN balancing technique would be a better choice to implement any real-time backorder prediction system. Among all five different classical machine learning models, the gradient boosting (XGBoost) classifier achieved higher accuracy. On the other hand, in terms of the most effective and important evaluation metric, AUC, support vector machine performs better than other models. In addition, other classification models including decision tree, SVM, and KNN also achieved effective performance in backorder prediction except for Gaussian Naive Bayes.

The experimental setup for deep learning techniques can be classified based on the parameters. The experiments are conducted by training CNN-based models with different settings. Two types of CNN models are applied. One has max-pooling layers and the other does not. The models were trained using two different epoch sizes which are 50 and 100. The performance of all experimental settings is presented in Table 22.

Table 22: Performance of CNN-based models in terms of *accuracy* and *AUC*. The best result is in **bold**.

Predictive Model	Balancing Tech.	Accuracy	AUC
CNN_50	ADASYN	0.8756	0.9443
	SMOTE	0.8936	0.9425
CNN_100	ADASYN	0.8868	0.9460
	SMOTE	0.8938	0.9432
MxCNN_50	ADASYN	**0.8947**	0.9475
	SMOTE	0.8938	0.9432
MxCNN_100	ADASYN	0.8903	**0.9489**
	SMOTE	0.8877	0.9462

From the results, we can see that the convolutional neural network-based model with max-pooling layer (MxCNN_100 and MxCNN_50) performed better among other experimental settings. It can also be concluded that our introduced ADASYN data balancing technique achieved efficient performance in both evaluation metrics. We added dropout layers that randomly drop some neurons in the training process for regularisation to overcome the over-fitting problem. To illustrate the necessity of dropout layers in CNN-based models, we carried out experiments with and without dropout layers. The performance based on the evaluation metrics concludes that dropout layers overcome the over-fitting problem. The MxCNN model without dropout layers achieved accuracy and AUC in the training data of 0.9081 and 0.9651, respectively. On the other hand, for testing data, the performance is lower in terms of both metrics. Without dropout layers, the performance of the model on test data based on accuracy and AUC are 0.8792 and 0.9411, respectively. Though the performance difference (2.89% in accuracy and 2.4% in AUC) between the training and testing data is not that big but still it has over-fitting. The method with dropout layers got the training accuracy and AUC of 0.8843 and 0.9499, respectively. For test data, the performance is quite consistent with accuracy and AUC of 0.8903 and 0.9489, respectively. Thus, we can say that the inclusion of dropout layers overcomes the over-fitting problem and eventually increases the

performance.

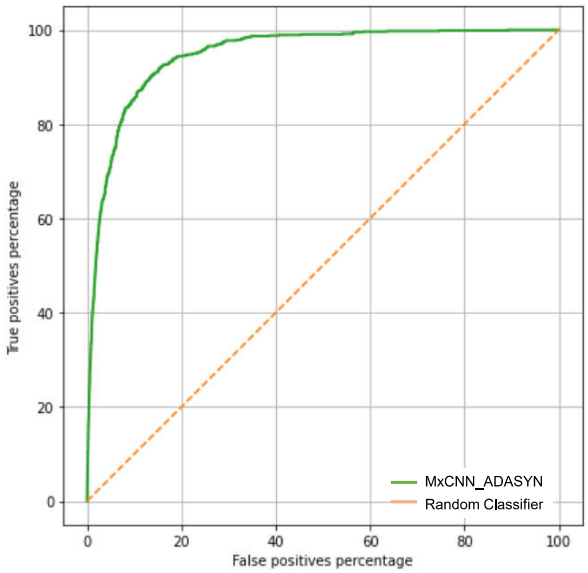

Figure 38: Performance of convolutional neural network based predictive model in terms of Receiver Operating Characteristic curve (ROC curve). The X-axis and Y-axis indicate the false positive and true positive percentage, respectively.

As compared to the performance of classical machine learning models reported in Table 21, the prediction power of our proposed CNN-based approach is way higher than the performance of ML methods. Although classical machine learning classifiers achieved higher *accuracy*, our CNN based model achieved a huge improvement in predicting future backorder prediction in terms of *AUC*, which is a more important metric to judge the performance of a classifier on data imbalance, because higher accuracy alone might not guarantee the predictive power of a classifier in case of extreme data imbalance. The performance of our method is also depicted by the Receiver Operating Characteristic curve (ROC curve) in Fig. 38. The curve illustrated the performance of our predictive model as compared to a random classifier. The area within the green curve shows the higher AUC achieved in predicting product backorders.

9.5.3 Performance Comparison with State-of-the-art Methods

The performance comparison of our proposed backorder prediction model with existing state-of-the-art methods is presented in Table 23. We directly reported the performance in terms of accuracy and AUC from existing published papers. Some existing works reported the performance only in terms of AUC but did not use accuracy and some others did the opposite. The blank cells (i.e., "-") in the table indicate that the performance based on particular evaluation metric is not reported in the published paper. According to the performance of different state-of-the-art methods reported in the table, we can see that our CNN-based predictive model outperformed the known related works in terms of both evaluation metrics except for one method by [286]. In turn, our methods significantly outperformed all methods based on accuracy. Shajalal et al. [286] applied a deep neural network with SMOTE oversampling technique. They applied four different variants of their methods utilising oversampling and under sampling techniques. Compared to the performance of those methods, our model got the best performance except one. Though one of their methods achieved a higher performance in terms of AUC, the performance difference with our method is subtle. In addition, their method lacks global interpretability and local explainability.

Table 23: Performance comparison of our method with known related work on the same dataset in terms of *accuracy* and *AUC*. The performance of our method is in **bold**.

Method	**Technique**	**Accuracy**	**AUC**
Our Method	ADASYN + CNN	**0.8947**	**0.9489**
Ntakolia et al. [226]	NN (MLP)	0.8568	0.9200
Shajalal et a. [286]	SMOTE_DNN	–	0.9586
	Weighted_DNN	–	0.9350
	Ran_Over_DNN	–	0.9475
Islam et al. [134]	DRF	0.8436	0.7870
	GMB	0.7919	0.7950
Hajek et al. [119]	RF	–	0.9157

[134] applied a distributed random forest (DRF) and gradient boosting machine (GBM) classifier to model product backorder. The performance of their models is struggling compared to the CNN-based model in terms of both evaluation metrics. Another noticeable concern in the performance of their method is substantial over-fitting. Numerically, their training accuracy of 0.9835 is way higher than the testing accuracy of 0.8436. Another work by [119] applied classical machine learning classifiers to model product backorder prediction. From their applied ML classifiers, random forest (RF) achieved the best AUC, which is still lower than our method. Note that they did not report the accuracy in the paper. Similar to Shajalal et al. [286], Ntakolia et al. [226] proposed a multi-layer perceptron (MLP) based neural network (NN) for modelling product backorder. But their performance is still much lower than ours in terms of both evaluation metrics. We think adding ADASYN oversampling technique overcome the data imbalance problem better. With this, our convolutional neural network-based predictive model capture the product backorder more efficiently as compared to other state-of-the-art methods. Having this comparative analysis, we can conclude that our method has got a new state-of-the-art performance in predicting product backorder in the inventory management system.

9.6 Explaining Backorder Prediction Model

This section presents the explainability of our introduced XAI techniques to interpret and/or explain the overall model and particular decisions of the proposed backorder prediction model. We first present the global model agnostic explanations generated to interpret the overall model's prediction priorities. Then the local explanations for a certain prediction are presented to provide the overall insight to understand a certain product will be going to be backordered or not.

9.6.1 Explaining overall model's priority

To interpret and explain the overall model, we exploit Shapely Additive values (Shap Values) that highlight the overall features' contributions in predicting the model's decisions. The feature contributions for the best performing model are depicted in Fig. 39 and 40. We can see both

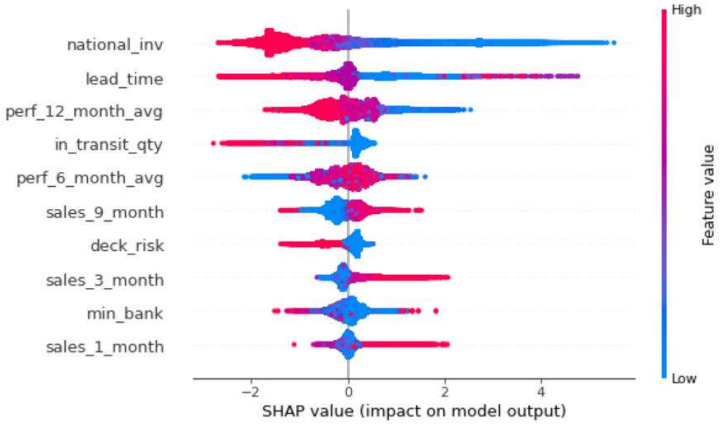

Figure 39: Global interpretation of the features' contributions of backorder prediction model as summary plot.

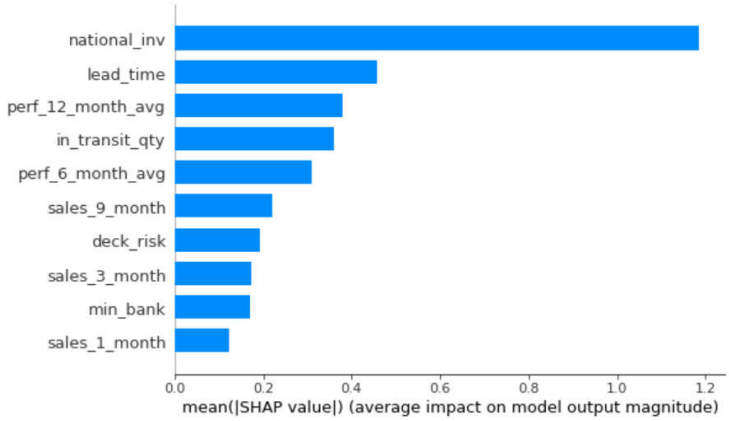

Figure 40: Global interpretation of the features' contributions of backorder prediction model as bar chart.

figures indicate the top ten most important features that the prediction model has given higher importance, such as current inventory, transit quantity, lead time, performance in the last 12 and 6 months, and sales.

We can conclude that these are the top 10 most important features on which the model depends more to predict whether a product is going to be backordered or not.

9.6.2 Explaining individual predictions

The features described in the previous figures (Fig. 39 and 40) have overall high importance in the predictive models. But it is expected that every sample (order) is different and unique in terms of features' values. Therefore, the importance and contributions of different features also will be different for particular order. To identify the most contributing features of each order, we employed local interpretable model agnostic explanation (LIME) to explain individual predictions. Using LIME, we trained a surrogate model with a portion of training data that mimics the performance and decision-making priorities of the proposed backorder prediction model. The explanation using LIME is depicted in Fig 41 and 42. These are the explanations for two individual backordered samples.

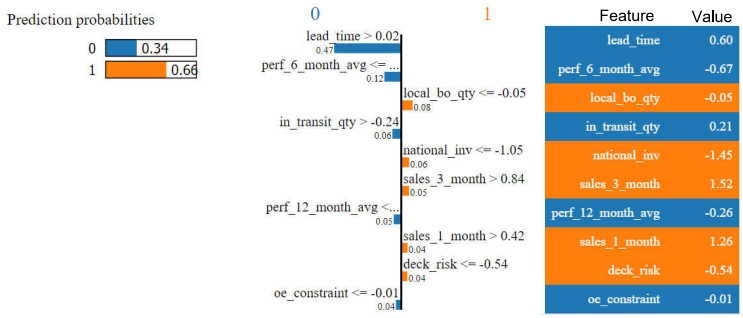

Figure 41: Local explanations of an individual prediction using LIME

The labels of these two products are 1 (backordered) and the model also predicted the same. The features in the right side marked by Yellow color pushed the predictive model to classifly as backordered, while Blue colored features in the left side did the opposite. From Fig. 41, we can

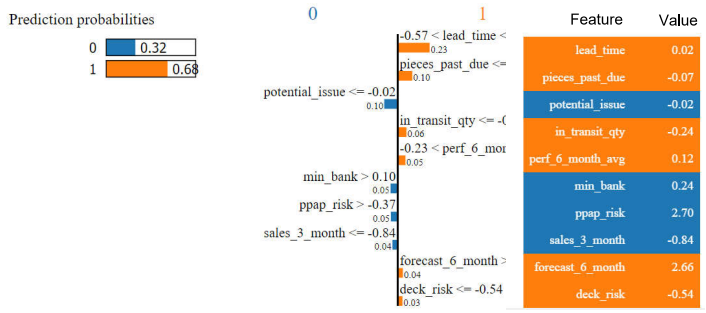

Figure 42: Local explanations of an individual prediction using LIME

see that the probability for being classified as backordered and non-backordered is 66% and 34%, respectively. The figure also indicates that the most important features (features in the right side) that lead to the prediction as backordered are local_bo_qty, current_inventory, and sales in the last 1 and 3 months and a risk flag. On the other hand, features (features in the left side) like lead time, in_transit_quantity, performance in the last 6 and 12 months, etc. try to push the model to predict a product as non-backordered. However, for another backordered sample, we can see that the list of contributed features for the backorder decision is different than the previous one. In Fig. 42, features such as lead time contributed the most to pushing the model to decide as a backordered one, which was the opposite for the previous sample.

Figure 43: Local explanations of an individual prediction using Force plot

To explain the local individual predictions more transparently, we applied shap values to plot the explanation as a force plot. Fig.43 and 44 illustrate the explanations for two different backordered samples. In both figures, the predictions from the models, referred to as base values are **0.67** and **0.75**, respectively for both samples. The closer the

Figure 44: Local explanations of an individual prediction using Force plot

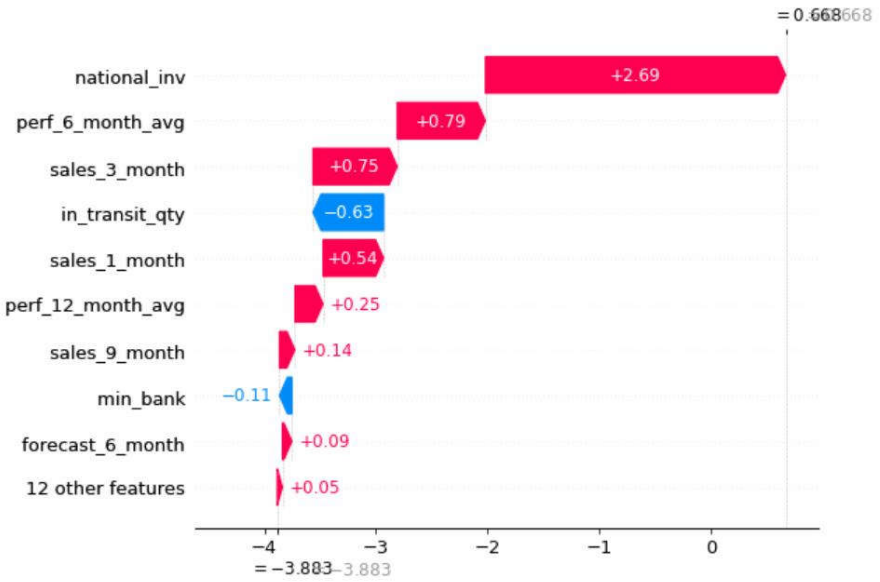

Figure 45: Local explanations of an individual prediction using Waterfall plot

value is to 1, the more the prediction leans toward backordered, while the closer to 0, the decision will then predicted as non-backordered. The red marked features contributed to increase the base value that help to decide the sample as a backordered one, and blue marked features did the opposite. The features having more impact on the base value remain closer to the boundary. For example, the two most-contributing features that push the model to decide the samples as backordered are current inventory and per_6_months for the first force plot (Fig. 43). For another example, the features with the most impact are current inventory and the forecasting for 9 months (Fig. 44). The explanations for the same two samples' decisions are also presented using the waterfall plot in Fig. 45

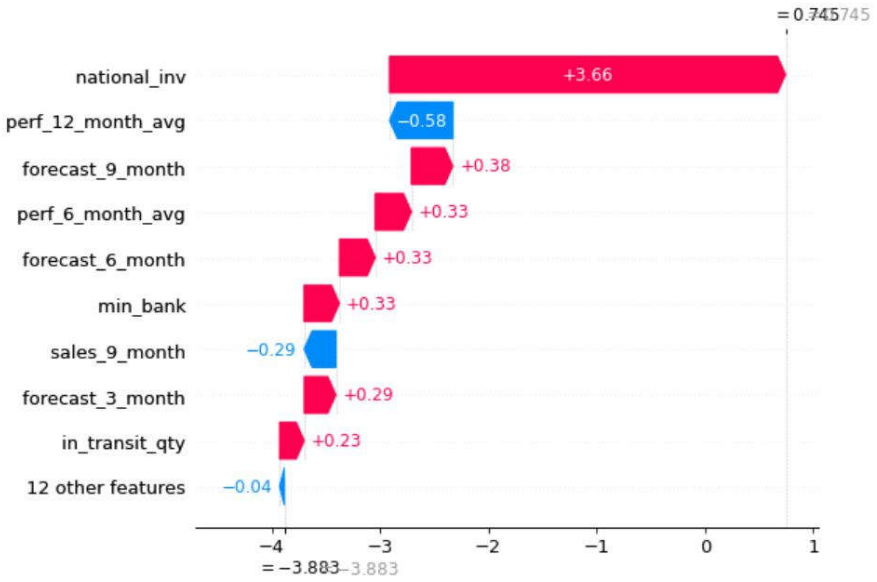

Figure 46: Local explanations of an individual prediction using Waterfall plot

and 46. The red marked features contributed to predicting the sample backordered and the blue colored try to push the classifier to predict the sample as non-backordered. Here the number and the span indicate the level of contributions of the features towards the decision.

With the help of our approach, stakeholders without in-depth knowledge of how backorder prediction systems work can have a better understanding both in terms of how the models generally factor in different types of data for making their decisions, as well as be enabled to analyse concrete decisions (that might seem counter intuititve or risky) in more depth than is possible with existing approaches. By applying such visualisations in practice, stakeholders would thus be enabled to enact suggestions from AI based systems more competently, and adapt their business strategies and decisions accordingly. This has the potential to both improve the usefulness, as well as also the willingness to adopt such systems in practice. While our introduced explainabilty has still to be evaluated with users, its is not trivial to implement such systems

in practice. In this regard, our paper contributes a demonstration of the applicability, and shows how such techniques can be implemented in a way that provides value to other stakeholders than developers of machine learning systems, to whose such applications are currently targeted.

9.7 Conclusion & Future Directions

This paper proposed a novel CNN-based model for product backorder prediction in an inventory management system and introduced global and local explainability that can explain the overall model decision-making priorities and answer the "*why*" question regarding any specific prediction. First, we proposed a novel convolutional neural network-based prediction model incorporating ADASYN oversampling technique to address data imbalance problem. The performance carrying out diverse experiment setups concluded that our proposed CNN-based backorder prediction model achieved a new state-of-the art result in product backorder prediction. In addition, the performance comparison with some known related methods demonstrated that our methods outperformed others in terms of multiple evaluation metrics. Secondly, our model is not only able to predict the backordered item but also can explain the reasons why the model predicts that a product is going to be backordered. For doing so, we utilised existing successful XAI techniques, SHAP and LIME, to explain the overall predictive model and individual decisions. Using global explanations, the stakeholder, and inventory managers can have an idea and understanding of how the overall model is making the decision. On the other hand, they can explicitly know and analyse why a certain product has a high chance to be backordered in the future, leveraging the explanations for their business decisions. Hence, they can identify which attributes have the most impact on a particular decision, and then react by adapting controllable

attributes (i.e. current inventory, lead time, etc.). Therefore, even when our approach still needs to be evaluated in practice, we believe these explanations can help the stakeholders to make their decision and minimize future losses. Most importantly, these explanations can increase the trust, transparency, and accountability of the AI-based predictive models in business problems, thus helping to overcome limitations of existing approaches that are more like black boxes for the users. While our study demonstrated the applicability of XAI techniques in the business domain on the concrete example of backorder predictions, there are multiple application areas such as customer churn prediction, customer behavior prediction, credit-worthiness assessment, fraud detection etc. where our explainable predictive model can be applied.

In the future, we plan to develop a collaborative interface to represent the explanations so that people can understand the decision-making more efficiently. We are also planning to introduce counterfactual explanations to provide a clear understanding about what are the possible actions she needs to take into account in the future.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 955422.

Part IV

Explainable Models in NLP

10 Introduction

In this part, we focus on investigating the explainability of NLP applications. To be diverse and explore how explanations can be changed in different application scenarios, we proposed two explainable models for patent classification and fake review identification. In the last two parts (part II & III), we conducted experiments on smart home applications with time series data and business applications with tabular data. In this part, we demonstrated explanations of textual data for two different applications.

Chapter 11 is about an explainable patent classification system, where we proposed a layer-wise relevance propagation-based explainable deep learning model. The proposed system can help patent experts classify certain patents into specific classes, explaining why the corresponding patent is classified to a particular class. Applying heatmap and word cloud, we demonstrated the explanations in two different forms. It would be easier to understand the relevant scientific terms that contribute to the classification in those explanations.

With a similar LRP-based explainability technique, we proposed explainable fake review identification methods employing pre-trained transformer models in chapter 12. The outcome of the models can predict the fake review with high precision, with explanations highlighting the most contributed words. We also conducted an empirical evaluation of the generated explanations and found that the explanations can make sense to the general users to some extent. However, the evaluations also concluded that the explanations should consider the grammatical structure, sentence tone, and contributed words.

11 Unveiling the Black Box: Explainable Deep Learning Models for Patent Classification

The content of this chapter has been presented in the 1st World Conference of eXplainable Artificial Intelligence (xAI2023) which has been held in Lisbon, Portugal in July 2023 and the paper has been published in the proceedings of the conference as a full paper by Springer Nature. The information of the paper is given as follows:

Information of the Article: Md Shajalal, Sebastian Denef, Md. Rezaul Karim, Alexander Boden and Gunnar Stevens. 2023. Unveiling the Black Box: Explainable Deep Learning Models for Patent Classification. In *Proceedings of the 1st World Conference on eXplainable Artificial Intelligence 2023 (xAI2023)*. Communications in Computer and Information Science, Springer Nature Switzerland 457–474. https://doi.org/10.1007/978-3-031-44067-0_24 (*Reproduced with permission from Springer Nature*)

Abstract

Recent technological advancements have led to a large number of patents in a diverse range of domains, making it challenging for human experts to analyze and manage. State-of-the-art methods for multi-label patent classification rely on deep neural networks (DNNs), which are complex and often considered black-boxes due to their opaque decision-making processes. In this paper, we propose a novel deep explainable patent classification framework by introducing layer-wise relevance propagation (LRP) to provide human-understandable explanations for predictions. We train several DNN models, including Bi-LSTM, CNN, and CNN-BiLSTM, and propagate the predictions backward from the output layer up to the input layer of the model to identify the relevance of words for individual predictions. Considering the relevance score, we then generate explanations by visualizing relevant words for the predicted patent class. Experimental results on two datasets comprising two-million patent texts demonstrate high performance in terms of various evaluation measures. The explanations generated for each prediction highlight important relevant words that align with the predicted class, making the prediction more understandable. Explainable systems have the potential to facilitate the adoption of complex AI-enabled methods for patent classification in real-world applications.

Keywords

Patent Classification, Explainability, Layer-wise relevance propagation, Deep Learning, Interpretability.

11.1 Introduction

Patent classification is an important task in the field of intellectual property management, involving the categorization of patents into dif-

ferent categories based on their technical contents [167]. Traditional approaches to patent classification have relied on manual categorization by experts, which can be time-consuming and subjective [183]. However, due to the exponential growth of patent applications in recent times, it has become increasingly challenging for human experts to classify patents. The international patent classification (IPC) system, which consists of 645 labels for the general classes and over 67,000 labels for the sub-groups, reflects the magnitude of challenges in multi-level patent classification tasks [167]. Furthermore, patent texts are generally lengthy and contain irregular scientific terms, making them a challenging field of application for text classification approaches, as patents often include highly technical and scientific terms that are not commonly used in everyday language, and authors often use jargon to make their patents unique and innovative [178]. These factors contribute to the significant challenges associated with patent classification.

However, recent advancements in machine learning (ML) and deep neural network (DNN) have made significant progress in automating the patent classification process. In the past, classical ML models, such as support vector machine (SVM), K-nearest neighbour, and naive bayes, have been widely used to automatically classify patent texts [82]. However, more recently, several DNN models have been proposed to address the challenges associated with patent classification. Generally, these models represent patent text using word embedding and transformer-based pre-trained models [196, 141, 167, 183, 63]. The DNN models, including recurrent neural networks (RNN) and their variants such as convolutional neural networks (CNN), long short-term memory networks (LSTM), bidirectional LSTM (Bi-LSTM), and gated recurrent unit (GRU), can learn to classify patents based on their textual content [196, 63, 183, 97, 118]. Hence, these enable faster and more reliable categorization of patents and scientific articles.

Mathematically, DNN-based classification approaches are often complex in their architecture, and the decision-making procedures can be opaque [293, 195]. While these approaches may exhibit efficient performance in classifying patents, the decisions they make are often not understandable to patent experts, or even to practitioners of artificial intelligence (AI). As a result, it is crucial to ensure that the methods and decision-making procedures used in patent classification are transparent and trustworthy, with clear explanations provided for the reasons behind each prediction. This is particularly important because patents are legal documents, and it is essential to comprehend the reasoning behind the classification decisions made by the model. Therefore, patent classification models should be designed to be explainable, allowing the reasons and priorities behind each prediction to be presented to users. This will help build trust in the predictive models and promote transparency among users and stakeholders.

For text-based uni-modal patent classification tasks, explanations can be provided by highlighting relevant words and their relevance to the prediction, thus increasing trust of users in the accuracy of predictions. In recent years, there has been a growing interest in developing explainable artificial intelligence (XAI) to unveil the black-box decision-making process of DNN models in diverse fields, including image processing [28], text processing, finance [170, 277], and health applications [332, 5]. These XAI models can provide insights into the decision-making process, explaining the reasoning behind specific predictions, the overall model's priorities in decision making, and thereby enhancing the transparency and trustworthiness of the application [195, 244, 293, 42, 28].

In this paper, our goal is to develop a patent classification framework that not only predicts the classes of patents but also provides explanations for the predicted classes. To achieve this, we propose a new explainable method for patent classification based on layer-wise rele-

vance propagation (LRP). This method can break down the contribution of patent terms that are crucial in classifying a given patent into a certain class. We start by representing the patent terms using a high-dimensional distributed semantic feature vector obtained from pre-trained word-embedding models. Next, we proceed to train several DNN-based models, including Bi-LSTM, CNN, and CNN-BiLSTM, which are capable of predicting the patent class. Finally, the LRP-enabled explanations interface highlights relevant words that contributed to the final prediction, providing an explanation for the model's decision.

We conducted experiments using two benchmark patent classification datasets, and the experimental results demonstrated the effectiveness of our approach in both classifying patent documents and providing explanations for the predictions. Our contributions in this paper are twofold:

1. We propose an LRP-based explainability method that generates explanations for predictions by highlighting relevant patent terms that support the predicted class.

2. Our developed DNN models show effective performance in terms of multiple evaluation metrics on two different benchmark datasets, and performance comparison with existing works confirms their consistency and effectiveness.

Overall, explainable DNN models offer promising solutions for patent classification, enabling faster and more accurate categorization while providing insights into the decision-making process. With the increasing volume of patent applications, the development of such explainable models could be beneficial in automatically categorizing patents with efficiency and transparency.

The rest of the paper is structured as follows: section 11.2 presents the summary of existing research on patent classification. Our proposed explainable deep patent classification framework is presented in

section 11.3. We demonstrate the effectiveness of our methods in classifying patents and explaining the predictions in detail in section 11.4. Finally, section 11.5 concluded our findings with some future directions in explainable patent classification research.

11.2 Related Work

In recent years, the patent classification task has gained significant attention in the field of natural language processing (NLP) research, as evidenced by several notable studies [288, 183, 178]. Various methods have been employed for classifying and analyzing patent data, and the methods can be categorized based on different factors such as the techniques utilized, the tasks' objectives (e.g., multi-class or multi-level classification), and the type of resources used to represent the patent data (i.e., uni-modal vs multi- modal) [252, 63, 118]. However, traditional approaches have relied on classical ML and bag-of-words (BoW)-based text representation, which have limitations in capturing semantic and contextual information of the text, as they can only capture lexical information. With the advent of different word-embedding techniques such as *word2vec* by Mikolov et al. [175, 203], *Glove* by Pennington et al. [233], and *FastText* by Bojanowski et al. [46], the NLP research has been revolutionized with the ability to represent text using high-dimensional semantic vector representations [274, 275, 276]. More recently, there has been a growing trend in employing transformer-based pre-trained models, including deep bidirectional transformer (BERT) [81], robust optimized BERT (RoBERTa) [192], distilled BERT (DistilBERT) [261], and XLNet [333], for text representation in NLP tasks.

Shaobo et al. [183] introduced a deep patent classification framework that utilized convolutional neural networks (CNNs). They started by representing the text of patents, which was extracted from the title and abstract of the USPTO-2 patent collection, using a skip-gram-based word-

embedding model [183]. They then used the resulting high-dimensional semantic representations to train CNN model. Similarly, Lee et al. [178] also employed a CNN-based neural network model, however, they fine-tuned a pre-trained BERT model for text representations. A DNN-based framework employing Bi-LSTM-CRF and Bi-GRU-HAN models has been introduced to extract semantic information from patents' texts [63].

A multi-level classification framework [118] has been proposed utilizing fine-tuned transformer-based pre-trained models, such as BERT, XL-Net, RoBERTa, and ELECTRA[68]. Their findings revealed that XLNet outperformed the baseline models in terms of classification accuracy. In another study, Roudsari et al. [252] addressed multi-level (sub-group level) patent classification tasks by fine-tuning a DistilBERT model for representing patent texts. Jiang et al. [141] presented a multi-modal technical document classification technique called *TechDoc*, which incorporated NLP techniques, such as word-embedding, for extracting textual features and descriptive images to capture information for technical documents. They modelled the classification task using CNNs, RNNs, and Graph neural networks (GNNs). Additionally, Kang et al. [148] employed a multi-modal embedding approach for searching patent documents.

A patent classification method called *Patent2vec* has been introduced, which leverages multi-view patent graph analysis to capture low-dimensional representations of patent texts [97]. Pujari et al. [236] proposed a transformer-based multi-task model (TMM) for hierarchical patent classification, and their experimental results showed higher precision and recall compared to existing non-neural and neural methods. They also proposed a method to evaluate neural multi-field document representations for patent text classification. Similarly, Aroyehun et al. [18] introduced a hierarchical transfer and multi-task learning approach for patent classification, following a similar methodology. Roudsari et al. [253] compared different word-embedding methods for patent

classification performance. Li et al. [182] proposed a contrastive learning framework called *CoPatE* for patent embedding, aimed at capturing high-level semantics for very large-scale patents to be classified. An automated ensemble learning-based framework for single-level patent classification is introduced by Kamateri et al. [146] .

However, to the best of our knowledge, none of the existing patent classification methods are explainable. Given the complexity of the multi-level classification task, it is crucial for users and patent experts to understand the reasoning behind the AI-enabled method's predictions, as it classifies patents into one of more than 67,000 classes (including subgroup classes). Therefore, the aim of this paper is to generate explanations that highlight relevant words, helping users understand the rationale behind the model's predictions. Taking inspiration from the effectiveness and interpretability of layer-wise relevance propagation (LRP) in other short-text classification tasks [19, 20, 149], we have adopted LRP [28] as the method for explaining the complex neural networks-based patent classification model.

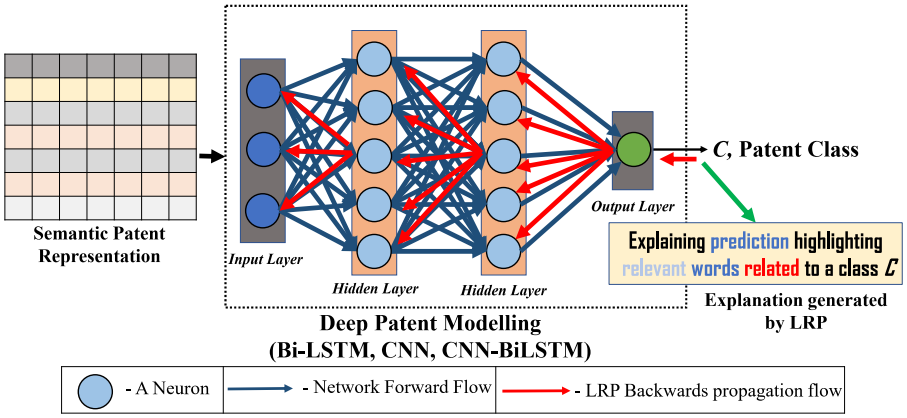

Figure 47: A conceptual overview diagram of our explainable patent classification framework.

11.3 Explainable Patent Classification

Our proposed explainable patent classification framework consists of two major components, i) training DNN-based classification model using the semantic representation of patent text, and ii) explanation generation component leveraging layer-wise relevance propagation (LRP). The conceptual diagram with major components is depicted in Fig 47. Our method first represents preprocessed patent texts semantically by high-dimensional vector leveraging pre-trained word embedding models. Then, the semantic representations for patent text are fed to train multiple DNN-based classification models including Bi-LSTM, CNN, and CNN-BiLSTM. For a particular deep patent classification model, our introduced LRP algorithm computes the relevance score towards a certain class for a given patent by redistributing the relevance score with backward propagation from the output layer to the input layer. Eventually, we get the score for patent terms that highlight the relevancy related to the predicted class of a given input patent.

11.3.1 Training deep neural models

Before training any specific DNN-based patent classification model, we employ *FastText* word-embedding model to represent each word of patent text with a high-dimensional feature vector and the element of each vector carries semantic and contextual information of that word. *FastText* is a character n-gram-based embedding technique. Unlike, *Glove* and *Word2Vec*, it can provide a word vector for out-of-vocabulary (OOV) words. Patents' text contains less used scientific terms and some words that are higly context specific. For example, patent in the field of chemistry has a lot of reagents and chemical names, even for some new patents the reagents' names might be completely new, proposed by the inventors. Considering this intuition, we chose *FastText* embedding in-

stead of *Glove* and *word2vec*. We make a sequence of embedding of the words for each patent and then fed it into the deep-learning model. Our trained different neural network models includes bidirectional LSTM (Bi-LSTM), convolutional neural networks (CNN), CNN-BiLSTM, a combination of CNN and Bi-LSTM.

11.3.2 Explaining predictions with LRP

Let c denotes the predicted class for the input patent p. The LRP algorithm applies the layer-wise conservation principle to calculate the relevance score for features. The computation starts from the output layer and then redistributes the relevance weight, eventually back-propagating it to the input layers [20, 19]. In other words, the relevance score is computed at each layer of the DNN model. Following a specific rule, the relevance score is attributed from lower-layer neurons to higher-layer neurons, and each immediate-layer neuron is assigned a relevance score up to the input layers, based on this rule.

The flow of propagation for computing the relevance is depicted by the red arrow that goes from the output towards the input layers in Fig. 48. The figure conceptually reflects how the semantic representation of patent text leads to a particular class in DNN models and back-propagates the relevance score from the output layer to the input layer for explanations highlighting relevant terms aligned with the predicted class.

The prediction score, $f_c(p)$ by our deep patent classification model, which is a scalar value corresponding to the patent class c. Using LRP, our aim is to identify the relevance score for each dimension d of a given patent vector p for the target patent class c. Our objective is to compute the relevance score of each input feature (i.e., words) that illustrates how positively (or negatively) contributes to classifying the patent as class c

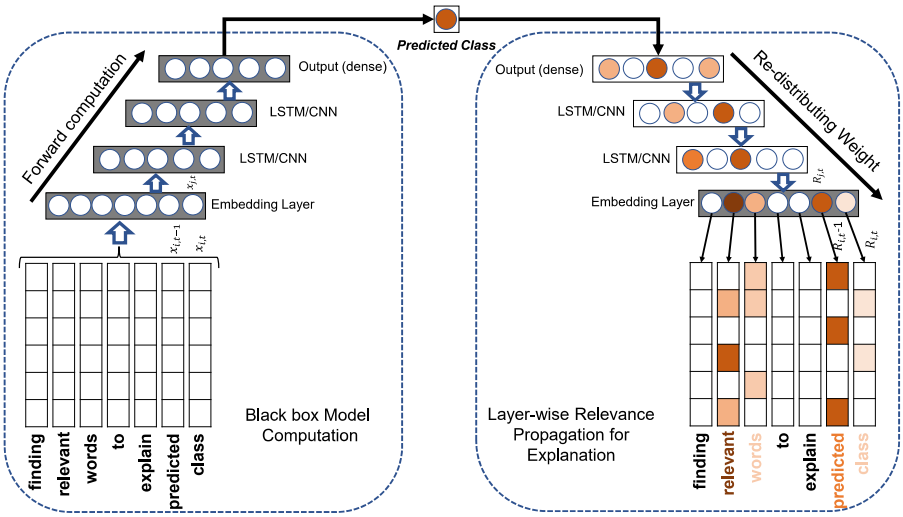

Figure 48: A conceptual overview diagram illustrating the working flow of layer-wise relevance propagation (LRP) (Figure created based on [19]).

(or another class).

Let z_j be the neuron of the upper layer and the computation of the neuron is calculated as

$$z_j = \sum_i z_i \cdot w_{ij} + b_j, \tag{11.1}$$

where w_{ij} be the weight matrix and b_j denotes the bias [20]. Given that the relevance score for upper-layer neurons z_j is R_j and we move towards lower-layer neurons to distribute that relevance. In the final layer, there is only one neuron (i.e., the prediction score) and in that case, R_j is the prediction score by the function $f_c(p)$. The redistribution of the relevance to the lower layers is done by following two major steps. We need to compute relevance messages to go from upper-layer to lower-layer neurons [20].

Let i be the immediate lower layer and its neurons are denoted by z_i. Computationally, the relevance massages $R_{i \leftarrow j}$ can be computed as followings [20].

$$R_{i \leftarrow j} = \frac{z_i \cdot w_{ij} + \frac{\varepsilon \cdot sign(z_j) + \delta \cdot b_j}{N}}{z_j + \varepsilon \cdot sign(z_j)} \cdot R_j. \tag{11.2}$$

The total number of neurons in the layer i is denoted as N and ε is the stabilizer, a small positive real number (i.e., 0.001). By summing up all the relevance scores of the neuron in z_i in layer i, we can obtain the relevance in layer i, $R_i = \sum_i R_{i \leftarrow j}$. δ can be either 0 or 1 (we use $\delta = 1$) [20, 149]. With the relevance messages, we can calculate the amount of relevance that circulates from one layer's neuron to the next layer's neuron. However, the computation for relevance distribution in the fully connected layers is computed as $R_{j \to k} = \frac{z_{jk}}{\sum_j z_{jk}} R_k$ [19]. The value of the relevance score for each relevant term lies in [0,1]. The higher the score represents higher the relevancy of the terms towards the predicted class.

11.4 Experiments

This section presents the details about the datasets, experiment results, and discussion of generated explanation with LRP.

11.4.1 Dataset

AI-Growth-Lab patent dataset: We conducted experiments on a dataset containing 1.5 million patent claims annotated with patent class[8] [37]. According to the CPC patent system, the classification is hierarchical with multiple levels including section, class, subclass, and group. For example, there are 667 labels in the subclass level [37]. However, for a better understanding of the generated explanations and the reasons behind a prediction for a given patent, we modeled the patent classification task with 9 general classes including *Human necessities, Performing operations; transporting, Chemistry; metallurgy, Textiles; paper, Fixed constructions, Mechanical engineering; lighting; heating; weapons; blasting engines or pumps, Physics, Electricity* and *General*.

[8]Dataset: https://huggingface.co/AI-Growth-Lab

BigPatent dataset: BigPatent[9] dataset is prepared by processing 1.3 million patent texts [292]. However, the classification dataset contains in total of 35k patent texts with 9 above-mentioned classes as labels. They provided the dataset by splitting it into training, validation, and testing set, the number of samples are 25K, 5K, and 5K, respectively. There are two different texts for each patent, one is a raw text from patent claims and another version is the human-generated abstract summarized from the patent claims.

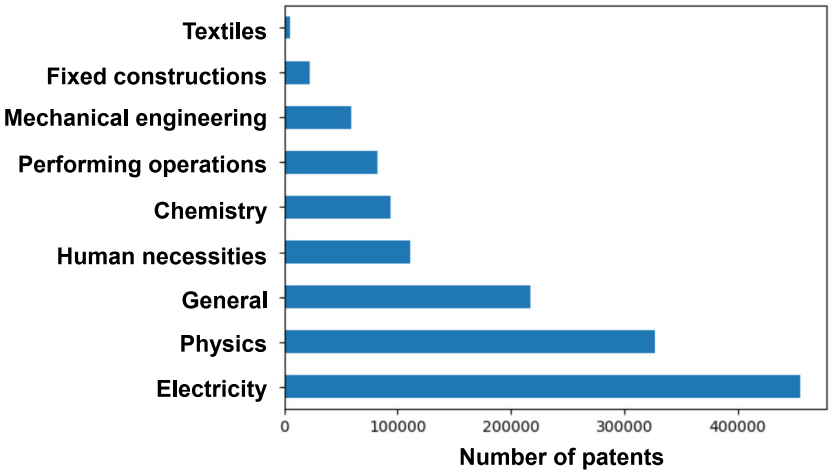

Figure 49: The distribution of the patents for different class on AI-growth-Lab data

However, the number of samples per patent class is varied widely for both both datasets, which means both are imbalanced dataset. The horizontal bar chart in Fig. 49 and 50 show the level of imbalance for both datasets. This imbalance distribution of samples per class poses an additional challenge in this multi-level classification task.

11.4.2 Experimental setup

We conducted experiments using three different DNN models, namely Bi-LSTM, CNN, and CNN-BiLSTM, utilizing the *FastText* pre-trained

[9]Dataset: https://huggingface.co/datasets/ccdv/patent-classification/tree/main

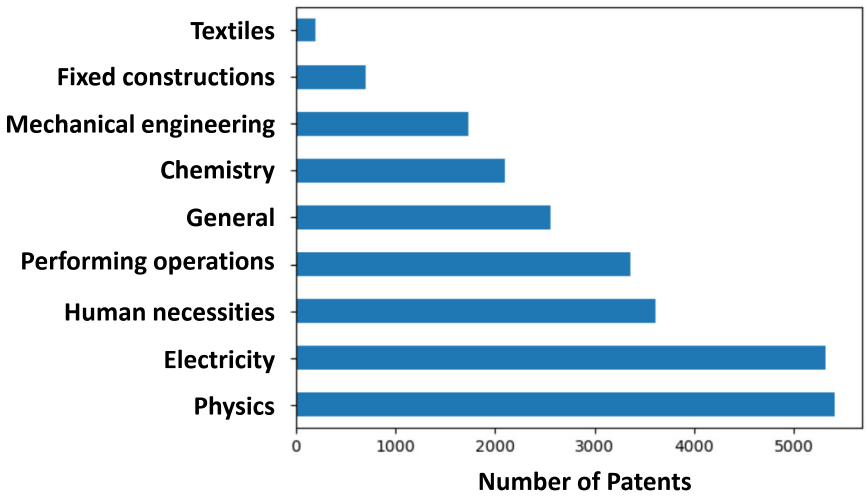

Figure 50: The distribution of the patents for different class on BigPatent data

word-embedding model for text representation in the embedding layers. The Bi-LSTM model consists of a layer of Bi-LSTM with 64 units after embedding layer, followed by another Bi-LSTM layer with 32 units, and then two fully-connected layers with 64 and 9 units, respectively. We applied the rectified linear units (ReLU) activation function in the hidden dense layer, and the softmax activation function in the output layer. For the CNN model, after the embedding layer, we have a 1-dimensional convolutional layer followed by a global average pooling layer, and finally, the output layer is a fully-connected layer with 9 units. The CNN-BiLSTM model has a convolutional layer followed by a global average pooling layer, and then the Bi-LSTM part is similar to the above-mentioned Bi-LSTM model. The activation functions in the fully connected hidden and output layers are ReLU and softmax, respectively. We implemented our methods using *scikit-learn* and *Keras*, and represented the patent text using the *FastText* pre-trained word-embedding model[10]. For implementing LRP for the Bi-LSTM network, we followed the method described in [20][11]. For the BigPatent dataset, the training, testing, and

[10]https://fasttext.cc/docs/en/crawl-vectors.html
[11]https://github.com/ArrasL/LRP_for_LSTM

validation sets are already split. For the AI-Growth-Lab data, the ratio for the training and testing set is 80% and 20%, respectively.

Table 24: The performance of different deep patent classification models on two datasets in terms of precision, recall and f1-score. The best result is in **bold**.

Dataset	Method	Precision	Recall	F1-Score
AI-Growth-Lab	**Bi-LSTM**	**0.69**	**0.70**	**0.69**
	CNN	0.62	0.63	0.62
	CNN-BiLSTM	**0.69**	0.68	**0.69**
BigPatent	**Bi-LSTM**	**0.79**	**0.78**	**0.78**
	CNN	0.75	0.76	0.76
	CNN-BiLSTM	0.77	0.76	0.76

11.4.3 Performance analysis

The performance of the proposed classification models was evaluated using three evaluation metrics, including Precision, Recall, and F1-Score, on two datasets, as shown in Table 24. The results demonstrate consistent performance across most of the deep classification models. Among them, the Bi-LSTM model exhibited better performance in terms of all evaluation metrics on both datasets. However, the performance of the other two models, CNN and CNN-BiLSTM, was also consistent and effective, though slightly lower than the Bi-LSTM model. Specifically, for the first dataset, CNN-BiLSTM performed equally well in terms of Precision (0.69) and F1-Score (0.69), while the performance of the CNN-based model was comparatively lower for the AI-Growth-Lab dataset, with a Precision of 62%, which was 7% lower than the best-performing Bi-LSTM model. However, for the BigPatent dataset, the CNN model exhibited considerably better performance, with a Precision of 75%, which was only 4% lower than the Bi-LSTM model. The performance difference between the models for the other two metrics was even lower, at 2%.

The performance of all DNN-based classifiers on the BigPatent dataset

Table 25: Class-wise performance of Bi-LSTM model on BigPatent Dataset

Patent Class	**label**	**Precision**	**Recall**	**F1-score**
Human necessities	1	0.79	0.91	0.85
Performing_operations	2	0.74	0.66	0.70
Chemistry	3	0.75	0.88	0.81
Textiles	4	0.71	0.74	0.73
Fixed_constructions	5	0.65	0.70	0.67
Mechanical_engineering	6	0.60	0.84	0.70
Physics	7	0.75	0.82	0.78
Electricity	8	0.78	0.86	0.82
General	9	0.71	0.46	0.41

is significantly superior compared to the first dataset. This may be attributed to the fact that the BigPatent dataset includes finely-grained abstracts of patents which are generated by human assessors, taking into consideration the patent texts. As a result, the semantic representation of the fine-tuned text in the BigPatent dataset is enriched compared to the raw patent claims in other dataset. We present the performance of Bi-LSTM model by showcasing the class-wise performance on the BigPatent dataset. Table 25 displays the performance across nine different patent classes. The Bi-LSTM model demonstrates favorable and consistent performance across most patent classes, with the exception of the *general* category. It is hypothesized that the patents in the *"general"* category may contain more commonly used terms compared to patents in other area-specific categories. Consequently, the captured semantic information may not be sufficient, potentially resulting in lower performance in terms of recall and F1-Score for the *"general"* class compared to other classes.

We compared the performance of our models with similar models that used *FastText* embedding for patent text representation. Compared to existing works by Roudsari et al. [118] and Shaobo et al. [183], the performance of our trained models is effective. Roudsari et al. also trained

Table 26: Performance comparison with related works

Method	Precision	Recall	F1-Score
Out Method	0.79	**0.78**	**0.78**
Roudsari et al. [118] (Bi-LSTM)	0.7825	0.6421	0.68.42
Roudsari et al. [118] (CNN-BiLSTM)	0.7930	0.6513	0.6938
Shaobo et al. [183] (DeepPatent)	**0.7977**	0.6552	0.6979

similar models with semantic text representation with a pre-trained *FastText* word-embedding model. They also develop similar DNN models including Bi-LSTM and CNN-BiLSTM. Shaobo et al. [183] introduced CNN-based deep patent modelling employing *FastText* word-embedding model. The performance of our methods on BigPatent data is higher than their models for all evaluation metrics except Precision. The comparison shows the effectiveness of our methods in classifying patents.

11.4.4 Generated explanation for prediction

We attempted to unbox the black-box nature of the deep patent classification model by adopting a layer-wise relevance propagation technique to compute the relevance score for each term by back-propagating the prediction score from the output layer to input layers. To represent the explanation per predicted class for a given patent text, we highlighted the related words that contributed to the classifier's prediction. As an example explanation, a patent is classified as *Chemistry*, and the related words that contributed to the prediction are highlighted in red color in Fig 51. The figure shows the explanation highlighting relevant words for the patent that classified as *chemistry*. The intensity of the color represents the contributions of a particular word. The higher the intensity of the color (red), the better the relevancy the word is. We can see that from the figure, the most relevant words include, *alkali, alkyl, monomer, acid, acrylate, acrylonitrile, acetate, polymer, ether*. We can observed that the highlighted words are completely related to terms

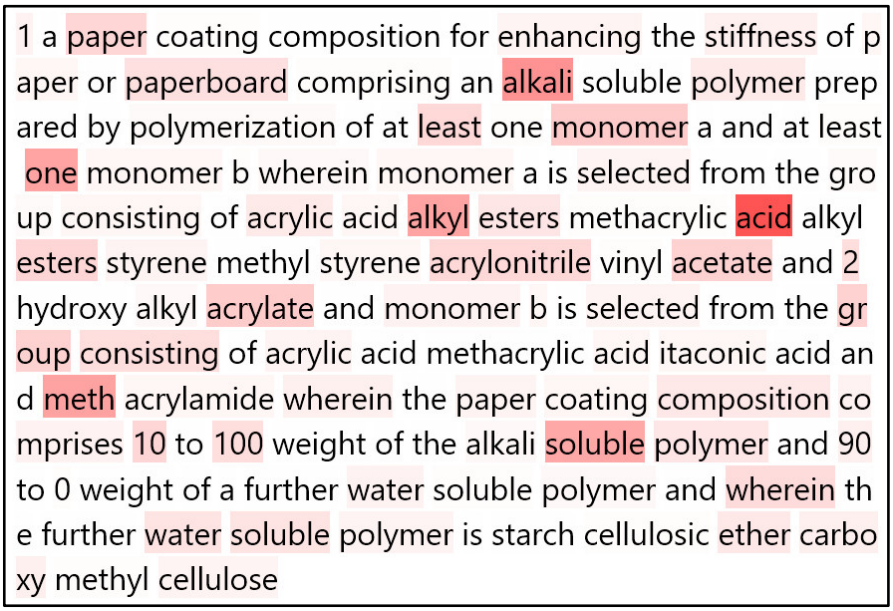

Figure 51: An example explanation for a patent classified as *Chemistry* patents highlighting relevant words. The higher the intensity of the color, the better the relevancy of the words contributing to the prediction.

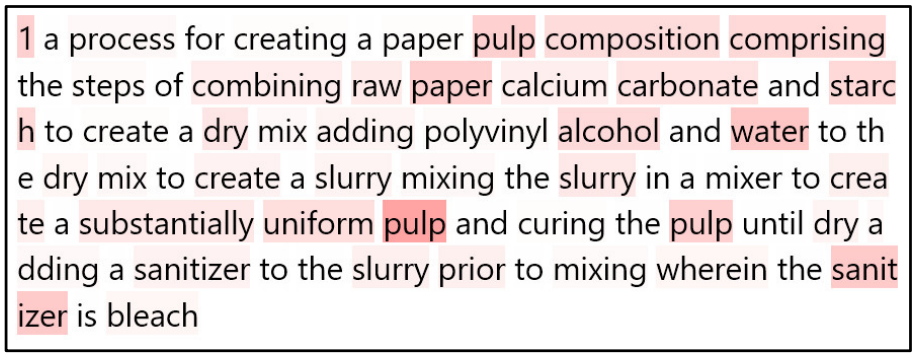

Figure 52: An example explanation for a patent classified as *Chemistry* patents highlighting relevant words. The higher the intensity of the color, the better the relevancy of the words contributing to the prediction.

used in organic chemistry and the explanation makes sense why this patent has been classified as a chemical patent. The next relevant list of words is *soluble*, *water*, *stiffness*, *enhancing*, *etc.* These words are directly related to chemistry except *stiffness* and *enhancing* . Since *enhancing* the *stiffness* of the paper or paperboard is the objective of this patent, these words are selected as relevant. Fig. 52 shows explanation

for another *Chemistry* patent. We can also see that the relevant words are highlighted including *pulp, alcohol, sanitizer, calcium, carbonate, etc.,* are directly related to the chemistry field. To understand the impact of particular words, we vidualize the words in the form of word cloud in Fig. 53. The larger the word, the higher the relevancy of the word to the certain class.

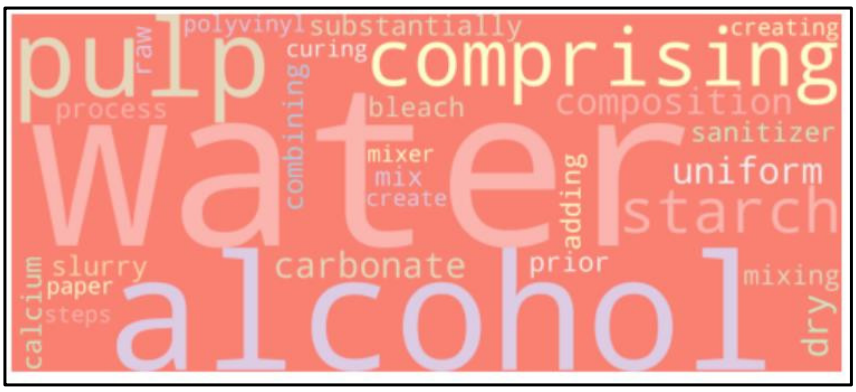

Figure 53: An example explanation for a patent classified as *Chemistry* patent highlighting relevant words in word cloud. The larger the font of the word, the better the relevancy of the words contributing to the prediction.

For another example patent in the field of *Electricity*, Fig. 54 illustrates the explanations highlighting relevant words that contributed to the classifier to decide that the patent is from *electricity* field. The most relevant words, in this case, include *power, channel, modem, device, bonded, bandwidth, data,* etc. We can see that all identified related words are used in *electricity* literature. The word *device* is used for common use in some other fields also, but this word also can be used to mention any electrical instrument in electricity-related explanation. Similar to the Fig. 53, we present the word cloud for explanation in Fig. 55. However, there are some words selected as relevant for both examples which are not relevant to the specific fields but can be used in literature for any field. One plausible reason is that those also might carry considerable importance in describing the any scientific object (i.e., explaining chem-

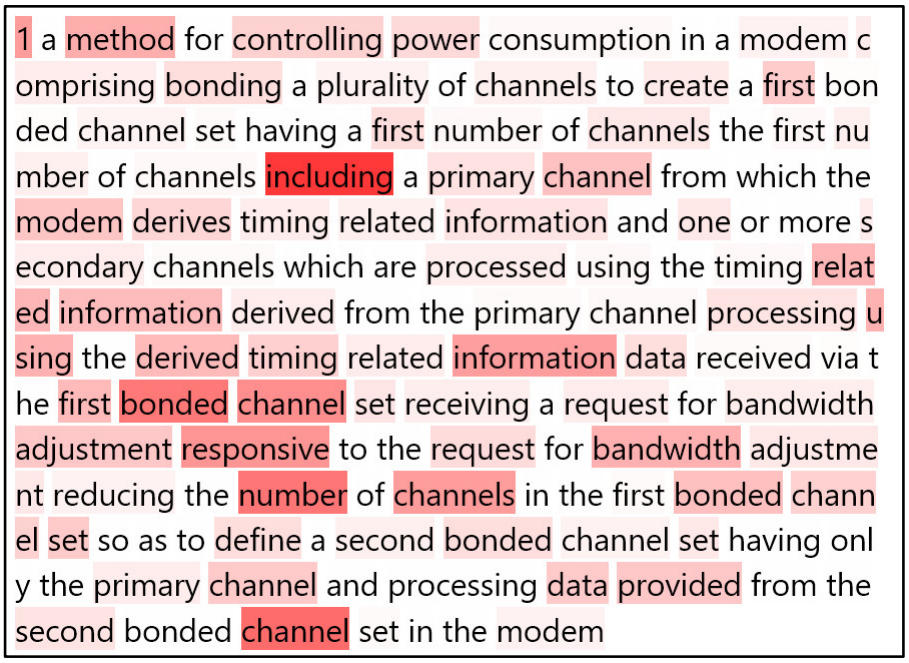

Figure 54: An example explanation for a patent classified as *Electricity* patents highlighting relevant words. The higher the intensity of the color, the better the relevancy of the words contributing to the prediction.

ical reaction) and capture good contextual and semantic information in *FastText* embedding.

Figure 55: An example explanation for a patent classified as *Electricity* patent highlighting relevant words in word cloud. The larger the font of the word, the better the relevancy of the words contributing to the prediction.

11.4.5 Limitations

Our model can explain the prediction for multi-label classification. Since the patents are classified in different levels and the patent classification system has a huge set of classes to classify in different levels, it should be explainable for multi-level classification also. This will be more challenging to explain the prediction for different subgroups-level classes. Another limitation is that our utilized pre-trained word-embedding model is not trained on the patent corpus. The local word-embedding model trained with patent corpus might capture better contextual and semantic information for scientific terms and jargon. Hence, the performance might be better than the current approach.

11.5 Conclusion and Future Direction

This paper aimed at explaining the predictions from DNN-based patent classification models with layer-wise relevance propagation technique to identify the relevance of different words in the patent texts for a certain predicted class. Layer-wise relevance propagation technique can capture context-specific explanatory and relevant words to explain the predictions behind certain predicted classes. The experimental results demonstrated the effectiveness of classifying patent documents with promising performance compared to existing works. We observed that the explanations generated by the LRP technique make it easier to understand why a certain patent is classified as a specific patent class. Most of the captured words have high relevancy with the patent domain, even though a few words marked as related are not that relevant (which, however, should also provide useful information to human expert in assessing the predictions). Even though our approach would still need to be evaluation with users, we can observe that the explanations are helpful to understand the question why a certain patent was classified into a

specific class, and to assess the results of deep-learning-based complex artificial intelligence-enabled models.

Since patents have a lot of scientific and uncommon words and phrases (i.e., jargon) that are not often used in other texts, we plan to train a local word-embedding model with patent texts to have better representation in our future work. It would be interesting to apply a transformer-based approach for the same purpose. The explanations for sub-group level prediction and capturing the sub-group context will be even more explanatory. However, the generated explanations will need to be evaluated by human experts in the patent industry. Therefore, we plan to have a user-centric evaluation for the generated explanations and elicit more human-centric requirements to be addressed in the future for better adoption real-word applications.

Acknowledgment

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 955422.

12 What Matters in Explanations: Towards Explainable Fake Review Detection Focusing on Transformers

The content of this chapter has been published as a preprint in ArXiv. The information of the paper is given as follows:

Information for Article: Md Shajalal, Md Atabuzzaman, Alexander Boden, Gunnar Stevens and Delong Du. 2024. What Matters in Explanations: Towards Explainable Fake Review Detection Focusing on Transformers, *Preprint in ArXiv*, 1-10. https://doi.org/10.48550/arXiv.2407.21056

Abstract

Customers' reviews and feedback play crucial role on electronic commerce (E-commerce) platforms like Amazon, Zalando, and eBay in influencing other customers' purchasing decisions. However, there is a prevailing concern that sellers often post fake or spam reviews to deceive potential customers and manipulate their opinions about a product. Over the past decade, there has been considerable interest in using machine learning (ML) and deep learning (DL) models to identify such fraudulent reviews. Unfortunately, the decisions made by complex ML and DL models - which often function as *black-boxes* - can be surprising and difficult for general users to comprehend. In this paper, we propose an explainable framework for detecting fake reviews with high precision in identifying fraudulent content with explanations and investigate what information matters most for explaining particular decisions by conducting empirical user evaluation. Initially, we develop fake review detection models using DL and transformer models including XLNet and DistilBERT. We then introduce layer-wise relevance propagation (LRP) technique for generating explanations that can map the contributions of words toward the predicted class. The experimental results on two benchmark fake review detection datasets demonstrate that our predictive models achieve state-of-the-art performance and outperform several existing methods. Furthermore, the empirical user evaluation of the generated explanations concludes which important information needs to be considered in generating explanations in the context of fake review identification.

Keywords

Fake Review, Explainability, LRP, Transformers, DistilBERT, XLNet, Empirical User Evaluation

12.1 Introduction

The rapid growth of e-commerce platforms for ordering various products online makes consumers' lives easier, saving potential time and cost for both ends. Issues related to trust and transparency are always of high importance, as they are directly associated with customer satisfaction and the revenue of companies or retailers [259]. Generally, customers or buyers in e-commerce or service providers tend to check the ratings and reviews of previous customers who have already purchased the products to get an idea of the quality of the targeted products. Users usually prefer to buy products with higher ratings and better reviews from customers. It is evident that companies or retailers sometimes take the opportunity to post fake positive reviews for their products with the objective of making their products appear better. Conversely, opposite scenarios are also visible, where competitors of certain products might post fake negative reviews to portray the products as being of poor quality.

However, identifying fake reviews can benefit both customers and retailers or companies by providing a trusted and transparent e-commerce platform. In the last decade, there has been a significant amount of attention on identifying fake reviews using automated methods with ML and DL-based classifiers. Notable ML methods such as SVM, NB, XGBoost, etc., generally use the TF-IDF or bag-of-words representation of textual reviews [205, 102]. However, these methods are traditional ways of representing text. The semantic representation of text using word embeddings has been employed in almost every natural language processing (NLP) task. Word embeddings can represent the semantic and contextual information of text in a high-dimensional space. With these representations, multiple methods have been proposed using DL-based classifiers, including recurrent neural networks (RNN) and its variants such as LSTM, BiLSTM, GRU, etc [285, 85, 335, 231, 230, 296, 80, 34]. After the invention of transformer-based text representations, NLP meth-

ods achieved high performance in almost every section. Transformer-based approaches, including BERT and its variants like DistilBERT, mBERT, and RoBERTa, have been used in many text classification tasks [149]. Recently, Electra [68], XLNet [333], GPT, and other large language models (LLMs) have also garnered significant attention in text classification, achieving high performance in numerous NLP tasks.

Generally, DL- and transformer-based approaches have complex architectures and involve a difficult decision-making process in predicting the original class. In the context of fake review detection task, users are typically laypeople with minimal knowledge about predictive models. The decisions might surprise them when they see a particular review detected as fake, but they cannot figure out why it is predicted as a fake review. Recently, explainable artificial intelligence (XAI) has gained significant attention in different fields, including business [277], bioinformatics [150], NLP [285, 149, 19, 20], and more. In the use case mentioned above, XAI comes into play to explain and validate the predictions made by the fake review detector. In this decade, XAI techniques have gained considerable attention in explaining model decisions, allowing AI practitioners and users to understand the reasons behind predictions and improve model performance and decision understanding. Several renowned XAI methods, including SHAP [195], LIME [244], LRP [20], Bert-interprete [241], can be applied to explain decisions related to NLP tasks.

The decisions made by complex ML models, which often function as "black boxes," can be surprising and difficult for general users to comprehend. In this research paper, we propose a transparent and efficient framework for detecting fake reviews, enabling high-precision identification of fraudulent content and providing users with explanations to help them understand the predictions. Initially, we develop fake review detection models using cutting-edge transformer models such as XLNet

and DistilBERT. We also applied different DL models, including BiLSTM, CNN, CNN-LSTM, and CNN-GRU models for detecting fake reviews. We then introduce LRP [19, 20] technique in fake review detection task to interpret the decisions from DL models and present explanations for individual predictions, highlighting the contributed words for the predicted class.

We conducted experiments in multiple settings, and experimental results on two benchmark fake review detection datasets demonstrate that our predictive models achieve state-of-the-art performance and outperform several existing methods. Furthermore, our generated explanations can interpret specific decisions, enabling users to understand why a particular review is classified as fake or genuine. The empirical evaluation with 12 human subjects was conducted to examine the effectiveness of the explanations and elicit further requirements in generating explanations in the context of fake review identification. The major contributions in this research are twofold:

- We introduced two transformer-based fake review detection models applying DistilBERT and XLNet that demonstrated significantly better performance than DL methods and existing related works.
- Our method is able to explain specific predictions with explanations introducing LRP techniques. The explanations might enable users to make sense of why particular reviews have been predicted as fake.
- Our conducted empirical evaluation of the generated explanations with human subjects and the results indicate further requirements in generating explanations for fake content identification tasks.

In the rest of the paper, we present the state-of-the-art methods in fake review detection in Section 12.2. The next section, Section 12.3, presents our method. The details of the dataset, experimental settings,

results on two different datasets, generated explanations with discussion and their empirical evaluation with human-subjects are presented in Section 12.4. Finally, we conclude our paper with some future directions in Section 12.5.

12.2 Literature Review

To understand people's strategies for creating and posting fake reviews, [29] conducted an empirical study with participants. Their findings suggested four stages involved in creating fake reviews: information gathering, assimilating information, drafting, and posting the reviews. These stages were identified through qualitative and quantitative methods with fifty-one participants. [169] highlighted two challenges in the theoretical grounding and under-researched areas of fake reviews. The first challenge is the lack of a conceptual understanding of the relationship between writing styles and recommendations. The second challenge is the knowledge gap regarding product characteristics. Their empirical investigation, which employed natural language processing (NLP) techniques, revealed latent characteristics of the product corresponding to buying preferences. However, their major findings suggested that the characteristics of fake reviews have no influence on recommending or discouraging the associated product. [234] analyzed the performance of different automated approaches for detecting deceptive customer reviews, with a focus on evaluating insights provided by multi-modal techniques.

Various classical and deep learning-based text classification models have been utilized, including support vector machines (SVM), k-nearest neighbors (KNN), logistic regression (LR), light gradient boosting model (LGBM) [254, 66, 90], long short-term memory network (LSTM), convolutional neural networks (CNN), recurrent neural networks (RNN), and transformers (e.g., BERT and its variants) for identifying the authen-

ticity of online reviews [85, 335, 231, 230, 296, 80]. [80] proposed an approach that combines CNN, particle swarm optimization (PSO), and various NLP techniques to identify the credibility and authenticity of online reviews using several datasets. Similarly, [85] presented a deep hybrid model for fake review identification, considering the combination of latent text features, aspect ratings, and overall ratings. [296] applied various classical ML and DL models for identifying fake reviews. [14] analyzed online deceptive reviews and proposed a recurrent neural network model for fake review detection, focusing on new features such as authenticity and analytical thinking. They conducted experiments on Amazon and Yelp reviews for electronic products.

Due to the hidden and diverse characteristics of fake reviews, developing a detection framework poses significant challenges. [294] proposed a detection framework based on the sentiment intensity of a review and positive unlabelled (PU) learning. [297] introduced an ensemble-based learning approach, which balances classes using different sampling techniques (WSEM-S), for modeling the fake review detection task. They applied n-gram models to extract features before applying the model. Their model's performance was compared with conventional ML models, including naive Bayes, XGBoost, KNN, and CNN. [205] applied different supervised models, including classical ML models, using BERT representations. [102] also employed supervised models.

A CNN-based fake news detection model was introduced by [314], which considers the web-scraped content heading. [339] proposed a deep learning-based method for fake news detection. [319] applied a voting-based approach to determine whether a review is fake or not, using multiple lists generated by different fake review detection models. [207] proposed an explainable fake review detection framework using different DL models, including Bi-LSTM, CNN, and DNN. They used Shapley Additive Explanations (SHAP) [195] to explain the models. Previously,

they addressed the concept drift problem within fake review detection systems [206, 208]. With the success of large language models (LLMs) in generating language for various purposes, it is evident that they are also frequently employed to generate artificial product reviews to influence customers' opinions [259]. [259] created a fake review dataset using multiple LMs and proposed detection methods for identifying fake reviews. [34] introduced an intelligent fake review detection method using different RNN variants, including CNN and LSTM. They extracted different aspects from the reviews and fed them into the deep learning models.

Topic modeling techniques were also applied by [43] to identify fake reviews. They combined review sentiment with other features for fake review identification. A semi-supervised Generative Adversarial Network (AspamGAN) model was proposed by [144], which incorporates an attention mechanism to address challenges related to the loss of important information due to the relative length of online texts. [313] introduced an ontology-based sentiment analysis approach that incorporates linguistic features and part-of-speech (POS) for identifying fake reviews. They employed a rule-based classifier based on different extracted features. Different handcrafted features have also been applied for identifying spam online reviews [290].Fang et al. [98] proposed a knowledge graph-based method that considers the time and semantic aspects of online customer reviews, as well as multi-source information. Feature-based and content-based classification methods have also been proposed [31, 54]. Content-based and product-related features have been applied for credibility detection in reviews [303].

From the above-detailed literature review analyzing published papers in the last five years, we can see that none of the methods are explainable except one [207]. But that one method only applied SHAP value-based explanation for global interpretability. Moreover, the methods proposed

to detect fake or deceptive or spam reviews lagged behind the current state-of-the-art methods including transformer-based methods. In this paper, we employ efficient transformers including XLNet and DistilBERT for modeling the fake review detection problem. We compared the performance of different models with baseline deep learning models including LSTM. BiLSTM, CNN, CNN-BiLSTM, and GRU using FastText word-embedding-based text representation. In addition, we applied layer-wise relevance propagation (LRP) based explanation technique to explain the prediction from our models.

12.3 Our Method

12.3.1 DistilBERT Transformer

DistilBERT [261] is a general purposed distilled version of BERT [81]. It is 40% lighter and 60% faster transformer model than BERT model. However, it retains 97% language understanding capabilities compared to BERT model. DistilBERT used knowledge distillation technique [124] which is a compression mechanism. In this mechanism a compact model, DistilBERT is trained to reproduce the behavior of large model, BERT. To reproduce the behavior of larger models, triple loss was employed in the training phase combining language modeling, distillation and cosine-distance losses. Besides, DistilBERT has the same architecture of the BERT model. As the number of layers have a smaller impact on the computation efficiency, the number of layers is reduced by a factor of 2 in the DistilBERT architecture. The token-type embeddings and pooler are also removed from the design of the compact model. Then DistilBERT is initialized from the BERT by taking one layer out of two and trained on the same corpus as the original BERT model. The technical details of DistilBERT can be found in the paper by [261].

12.3.2 XLNet Transformer

XLNet [333] is a generalized autoregressive approach that utilized the both of autoregressive and autoencoding language modeling. Unlike the usages of the autoregressive models' fixed forward or backward factorization order, XLNet maximizes the log likelihood of a text sequence with respect to all possible permutations of the factorization order to learn the bidirectional context. In contrast to BERT's pretrain-fineture discrepancy, XLNet does not suffer from it. It also use the predicted token's joint probability are factorized with the product rule provided by the autoregressive objectives. XLNet improved its performance by integrating the ideas of Transformer-XL [74] in its pretraining. As a result, XLNet outperforms BERT on 20 tasks including question answering, sentiment analysis, natural language inference and document ranking, etc. The technical detail of XLNet can be found in the paper by [333].

12.3.3 Explaining the prediction

Inspired by the success of the LRP technique in explaining text classification in different NLP applications, we adopted LRP [20, 19] to explain the prediction to answer "*why a particular review has been predicted as fake*?" question. The LRP technique can unveil the black-box deep learning model by back-propagating the output from the output layer up to the input layers through re-distributing the weight in the previous layers. In the end, LRP provides the weight for every input feature and the higher the weight the higher the relevancy of the words towards the predicted class. We then highlight the words based on the provided weight and represent the explanations in terms of highlighted text and word could.

For a deep neural network classification model, LRP re-distributed the weight from the output layers back to the input layers [285]. Let i be the

immediate lower layer and its neurons are denoted by z_i. Computationally the relevance messages $R_{i\leftarrow j}$ can be computed as followings [20].

$$R_{i\leftarrow j} = \frac{z_i \cdot w_{ij} + \frac{\varepsilon \cdot sign(z_j) + \delta \cdot b_j}{N}}{z_j + \varepsilon \cdot sign(z_j)} \cdot R_j. \tag{12.1}$$

The total number of neurons in the layer i is denoted as N and ε is the stabilizer, a small positive real number (i.e., 0.001). By summing up all the relevance scores of the neuron in z_i in layer i, we can obtain the relevance in layer i, $R_i = \sum_i R_{i\leftarrow j}$. δ can be either 0 or 1 (we use $\delta = 1$) [20, 149, 285].

12.4 Experiments

12.4.1 Dataset

Fake Review Dataset: This fake review dataset contains 40000 reviews in total. Among them, 50% reviews were originally written by humans (i.e., reviews collected from Amazon). The rest of the reviews are fake, generated by two different language models including ULMFit (Universal Language model Fine-tuning) and GPT-2 [259]. The reviews are for products from 10 different categories [259]. The categories are *Books, clothing, shoes and Jewelry, Electronics, Home and Kitchen, Kindle, Movies and TV, Pet Supplies, Sports and Outdoors, Tools and Home Improvements, and Toys and Games*.

Yelp Review Dataset: We conducted experiments with another fake review dataset named *Yelp Review Dataset*. Compared to the previous one, this dataset is quite big and consists of more than 682K reviews and the distribution is quite imbalanced. The dataset is accessible at Kaggle[12].

[12]https://www.kaggle.com/datasets/abidmeeraj/yelp-labelled-dataset/data

12.4.2 Experimental Settings

To assess the efficiency of our introduced fake review detection methods, we conducted a comprehensive set of experiments using two distinct fake review datasets. We initially applied an ensemble machine learning (ML) method, employing a majority voting technique on four different ML models: Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), and XGBoost. Subsequently, we explored four distinct deep learning models, namely BiLSTM, CNN, CNN-LSTM, and CNN-GRU. In the BiLSTM model, we incorporated an embedding layer followed by a Spatial dropout layer. This was followed by two bidirectional LSTM layers, each consisting of 64 and 32 BiLSTM units, along with a dropout layer after each. The model then incorporated a fully connected layer and concluded with an output layer featuring the *sigmoid* activation function.

Similarly, the CNN model began with an embedding layer, followed by a Spatial dropout layer. A convolutional layer was introduced, followed by a dropout layer, and finally, two fully connected layers. The CNN-LSTM model represented a hybrid combination of CNN and LSTM layers. It started with an embedding layer, followed by a CNN layer, an LSTM layer, and two fully connected layers. The architecture for the CNN-GRU model closely resembled that of the CNN-LSTM model, with the key distinction being the replacement of LSTM layers with GRU layers. For our transformer-based approaches, we employed DistilBERT and XLNet transformers for the purpose of detecting fake reviews. Since there is a high chance of vocabulary mismatch problem, we employed *FastText*, a character embedding-based pre-trained word embedding model, to represent the reviews semantically for deep learning models. For transformer models, we employed *distilbert-base-uncased* and *xlnet-base-cased* for DistilBERT- and XLNet-based classification models, respectively. For all experiments, we split both datasets into train, test,

and validation sets in 70%, 15%, and 15%, respectively.

Table 27: The performance of different methods compared to baselines on Fake Review Dataset.

Type	Model	Accuracy	Precision	F1Score
Baseline	EnsembleML	0.8425	0.9147	0.9014
Deep Learning	BiLSTM	0.9556	0.9750	0.9466
	CNN	0.9252	0.9268	0.9259
	CNN-LSTM	0.9486	0.9454	0.9493
	CNN-GRU	0.9476	0.9751	0.9466
Transformers	DistilBert	**0.9592**	**0.9906**	**0.9821**
	XLNet	0.9580	0.9887	0.9779

Table 28: The performance of the fake review detection method for different product categories.

Category	Accuracy	Precision	F1Score
Home	0.9557	0.9786	0.9689
Sports	0.9544	0.9750	0.9686
Electronics	0.9461	0.9807	0.9479
Movies	0.9499	0.9689	0.9631
Tools	0.9508	0.9707	**0.9711**
Pet	**0.9612**	**0.9833**	0.9705
Kindle	0.9461	0.9795	0.9508
Books	0.9439	0.9772	0.9459
Toys	0.9341	0.9708	0.9325
Clothing	0.9351	0.9802	0.9295

12.4.3 Experimental Results

Performance on Fake Review Dataset: The performance of different fake review detection models on Fake Review dataset [259] is presented in Table 27 in terms of multiple evaluation metrics. Among four different deep learning models, BiLSTM performs better in terms of accuracy (0.9556) and F1-Score (0.9466). We can also see that CNN-GRU performs equally compared to BiLSTM in terms of F1-Score and Precision which is almost the same. However, the other two DL models CNN and CNN-LSTM also achieved consistent and effective performance. In terms of all evaluation metrics, our proposed two transformer-based fake

review detection models achieved significantly higher accuracy (0.9592), precision (0.9906), and F1-Score (0.9821) among all employed models. The performance difference between XLNet and DistilBERT is not significant and it is only a 1% difference in terms of precision. DistilBERT achieved more than 4% performance gain in terms of F1-Score.

To illustrate the performance of our best method (DistilBERT) in identifying fake reviews, we present the performance across different categories. Table 28 presents the performance for reviews in different product categories. Identifying fake reviews for *pet supplies* category, DistilBERT achieved 96% accuracy and 98% precision compared to the performance on other categories. On the other hand, it achieved higher F1-Score of 0.9711, which is higher across all categories.

Table 29: Performance comparison with existing method on fake Review Dataset. The model *OpenAI, NBSVM and fakeRoBERTa* are proposed by [259].

Method	**Precision**	**Recall**	**F1Score**
OpenAI	0.83	0.82	0.82
NBSVM	0.95	0.95	0.95
fakeRoBERTa	0.97	0.97	0.97
DistilBert	**0.9906**	**0.9738**	**0.9821**
XLNet	0.9887	0.9703	0.9779

We compared the performance of our transformer-based models with state-of-the-art methods on the same dataset by [259]. They applied three different classification models to identify fake reviews. First, they trained support vector machine-enabled classifier with Naive Bayes Feature (NBSVM). Then, an OpenAI model is specifically developed and applied for fake review detection. The OpenAI model is based on the idea of Roboustly Optimized BERT Pretraining Approach (RoBERTa) with fine-tuning. Finally, inspired by the performance on OpenAI, they designed a RoBERTa-based customized model called *fakeRoBERTa* as their final model. The comparison presented in Table 29 shows that our both transformer models XLNet and DistilBERT achieved significantly better per-

formance in terms of precision and F1-score. In terms of recall, their performance is quite consistent. However, the performance compared to the results of a wide range of experiment settings and existing methods, our introduced DistilBERT model demonstrated a new state-of-the-art performance in fake review detection tasks.

Table 30: The performance of different methods compared to baselines on Yelp Review Dataset.

Type	**Model**	**Accuracy**	**Precision**	**F1Score**	**AUC**
Baseline	EnsembleML	0.7848	0.7795	0.8156	0.6271
Deep Learning	BiLSTM	0.8947	0.8985	0.9444	0.7236
	CNN	0.8961	0.8966	0.9451	0.7330
	CNN-LSTM	0.8842	0.9007	0.9380	0.6780
	CNN-GRU	0.8964	0.8978	0.9452	0.7249
Transformers	DistilBert	0.9235	**0.9326**	**0.959**5	0.7958
	XLNet	**0.9349**	0.9278	0.9654	**0.8044**

Performance on Yelp Review Dataset: We also carried out experiments on another dataset to demonstrate performance consistency. As we noted earlier the Yelp dataset is quite imbalanced and the number of majority samples is way larger than the minority samples, we measure the performance in terms of area under curve measure along with accuracy, precision, and F1-Score. We present the performance for the Yelp dataset in table 30. The table summarized that transformers-based classification models here also performed better than the deep learning models and ensemble ML model. Unlike the performance in the previous dataset, XLNet achieved higher accuracy, F1-Score and AUC compared to the DistilBERT-based classifier. But for the other measure, in terms of precision , DistilBERT performed better. However, the performance difference is not that big but compared to the deep learning-based methods, both DistilBERT and XLNet outperformed significantly with a way higher AUC. AUC is considered one of the best evaluation metrics to measure the performance when the dataset is imbalanced.

Overall, the performance on this dataset is lower than on the previous dataset. There are several probable reasons. One is the size of

the dataset, the Yelp dataset is significantly larger than the fake review dataset and reviews are written by human. However, in the Fake review dataset, the fake reviews are generated by the large language models (LLM). Additionally, the Yelp data is considerably imbalanced. Since the reviews are generated by LLM, the transformer-based classification models might recognize the review patterns better than the reviews written by humans. However, considering the performance of a wide range of experiments on two different datasets, we can conclude that DistilBERT and XLNet achieved new state-of-the-art results in identifying fake reviews, both for human and machine-generated fake reviews.

12.4.4 Explainability of the predictions

We present explanations provided by LRP technique using highlighted text and word cloud where the color intensity in highlighted text and size of the words *WordCloud* represent the degree of relevancy towards the class. We demonstrated explanations for two predictions from each dataset. The first considered review is for a book. It is a fake review generated by a transformer-based language model and our model also predicted it correctly, as fake. The review is as follows:

Review 1: *"First, let me say I'm an avid reader and this is a book that I read as a child. I had to read it before. I could have a chance to take it to college. I still enjoy reading it as a kid. This book is still one of my favorite books. I have read the book over and over again and it is a must read. I just can't put it down. The only reason I gave it five stars is because I want to read more about the characters. I liked the way they interacted with the kids. I loved their reactions to."*

The explanation is depicted in Fig. 56. We can see the highlighted words are related to the predicted class. The highlighted text and word

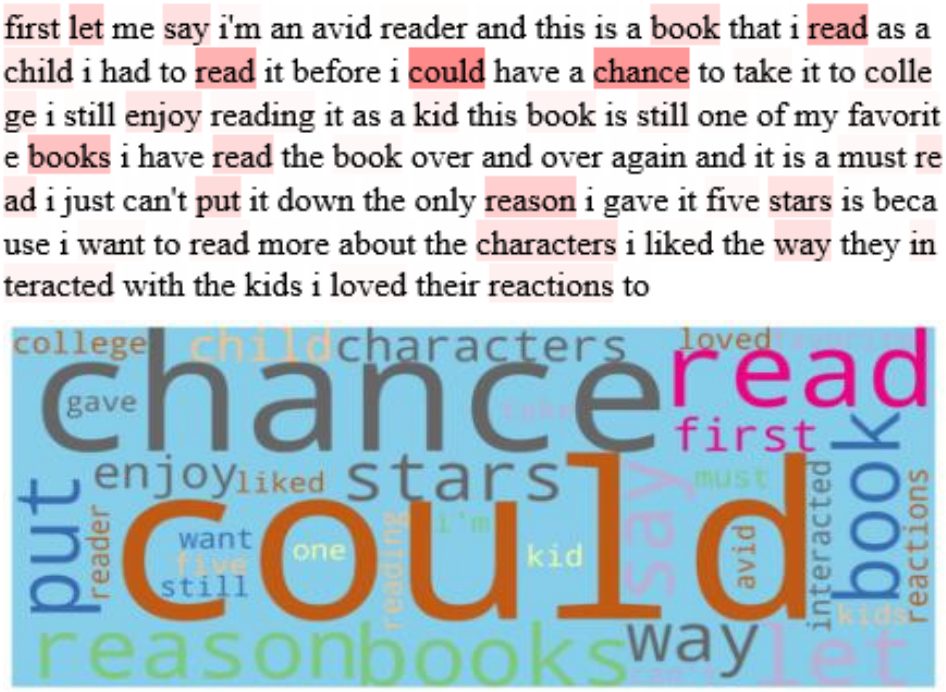

Figure 56: Explanation with highlighting relevant words for a predicted fake review.

cloud also show that words such as *read, chance, enjoy, liked, loved stars* are some most relevant for the prediction. We have a closer look at the review text, it is a review with exaggeratedly praises the book. The highlighted words are used for exaggerated praise. Let's consider another fake review for the clothing category:

Review 2: *"I actually have a review here on the site about these gloves. These gloves are awesome and i highly recommend them. I am a 32-year-old man who works at a large company and have had no issues with these gloves. I have had no issues with the gloves falling apart or breaking. They are the most comfortable I have ever worn, and I am so happy with them. I recommend them to anyone looking for a glove that will last a long time. I am very happy with my purchase and would recommend these gloves to anyone looking for a great glove. These are super lightweight and lightweight. I have used them for a couple of months, and they are*

still working great. I would definitely recommend them to anyone looking for a good quality product. I've had no problems with them slipping or sliding my only complaint is that I haven't used them yet, but they are very well made and have been used in my gym."

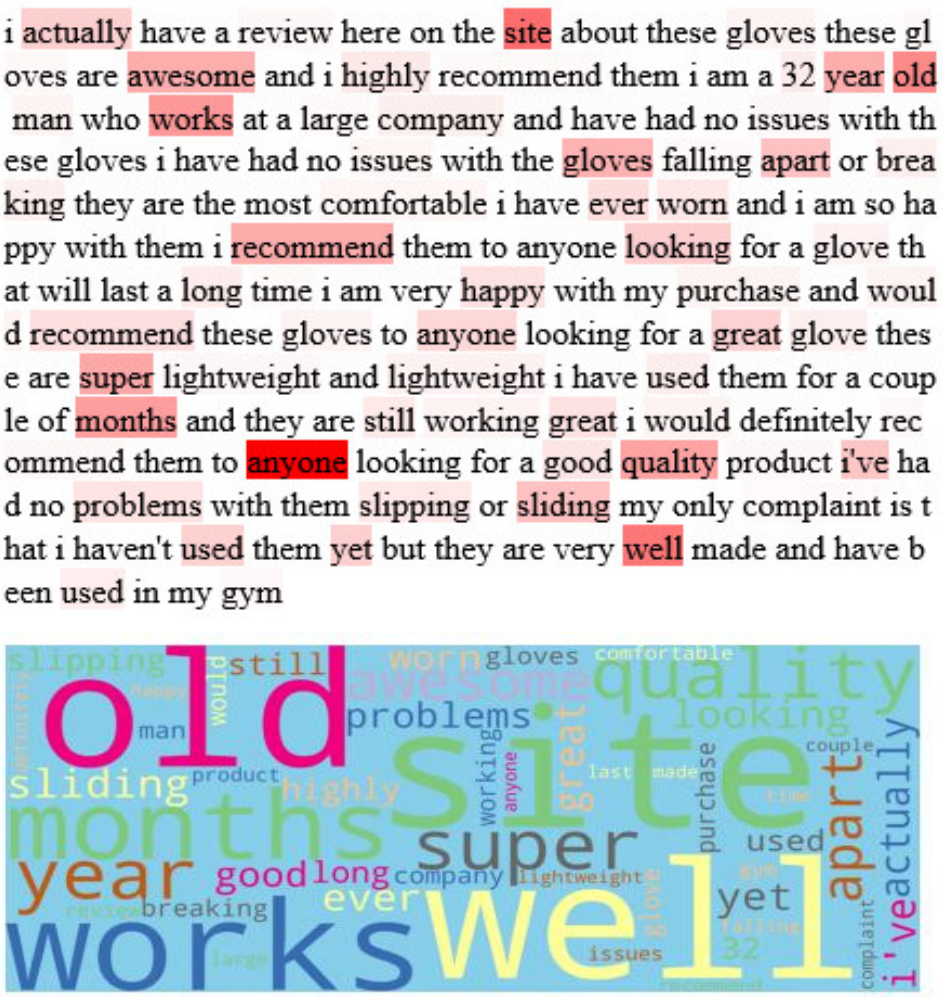

Figure 57: Explanation with highlighting relevant words for a predicted fake review.

The explanation for this review is illustrated in Fig. 57. This is a long review for *gloves* and if we have a closer look on the the text, we can see that it also has redundant praise (i.e., recommends the product several times and claims that the gloves were the best the user experienced ever). However, the LRP-enabled explanation identified words that in-

dicated why it is classified as fake. The relevant words are *awesome, recommend, anyone, quality*, etc.

In the Yelp dataset, the review text is not fine-grained since all the reviews, both real and fake, are written by humans. The grammatical quality is not similar to the previous dataset. However, let's look at the explanation of a review in Fig. 58. This is a review for a restaurant.

Review 3: *"Omg this place is highly recommended to me by a friend and I'm happy that I come here. It was fabulous. Everything was excellent. Amazing food and service thank you for everything David. Such a amazing service. You made my friend birthday great."*

Figure 58: Explanation with highlighting relevant words for a predicted fake review for Yelp Dataset.

We can see from the figure that the highlighted words including *amazing, fabulous, great, service, highly, recommended* etc. play bigger roles in helping the classifier to decide it is a fake review. For another review which is quite longer than the previous one. Fig. 59 presents that the responsible words for which the deep learning models decide it as a

fake review are *authentic, taste, favorites, dishes, open, place,* etc.

Review 4: *"NoodleTown is classic authentic chinese food. â The taste is always there prices are not bad maybe 50 cents or 1 more than other Chinatown restaurants but the food is good. Most of the dishes on the menu is good. â Our favorites are beef Chowfun, Chicken Chowmein in black pepper, sauce hoisin chowmein, seafood Congee with Cruellers if they didn't run out yet. Wonton Noodle, Soup dishes are good seafood. Dumpling is good this place accomodates until late in the night. They close around 4 am and re open soon after that definitely convenient and good food."*

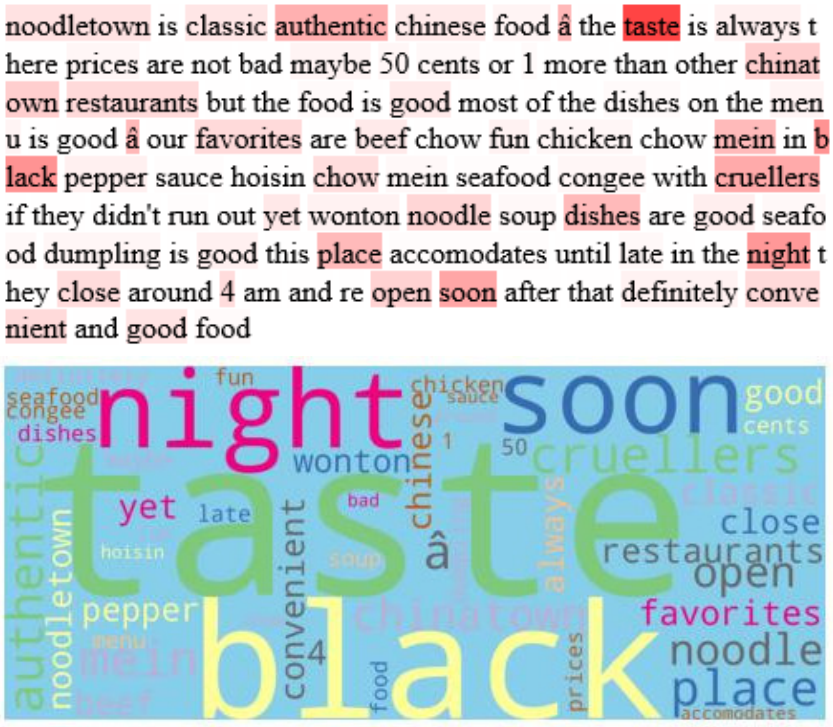

Figure 59: Explanation with highlighting relevant words for a predicted fake review for Yelp Dataset.

12.4.5 Empirical User Evaluation

To evaluate the effectiveness of the LRP-generated explanations highlighting the important relevant words to the predicted class, we conducted an empirical user study with 12 human subjects. The subjects are studying master's in business informatics. We first give then an overview of how our transformer-based model predicts the authenticity of the review. Then we provide them with a simple demo about the explanations and what those highlighted words mean.

We provided them three reviews (Review 1, Review 2, and Review 3) and asked them to score how authentic the reviews were. All three reviews were fake but we have not told them. Because we wanted to observe how they identify and what are the logic behind. We also provided the details about the products for which the reviews were posted. They were asked to put score for each review, and the score ranges from one star (*) to five star (*****). The highest value 5 (*****) indicates that the corresponding review is original, while the lowest value 1 (*) indicates the review is fake. The participants first score each review after carefully reading the reviews without the generated explanations.

We then provided them with the LRP-generated explanations (Fig., 56,57,58) for each reviews, respectively. We then instructed the participant to look at the explanations and re-score the reviews whether their assumptions changed after perceiving the explanations. We denote two scores before and after the explanations as *score 1* and *score 2*, respectively.

We then discuss with every participant why they think a particular review is original or fake. What are the reasons and rationale behind their scores? We also asked them about the efficiency of the generated explanations, and how they help the participants to decide on the authenticity of the reviews.

Table 31: The user evaluations whether the reviews are fake or real, with and without explanations.

Subject	**Review 1**		**Review 2**		**Review 3**	
	Score1	Score2	Score1	Score2	Score1	Score2
1	**	**	****	***	**	**
2	**	**	****	****	****	****
3	*	*	*	*	**	**
4	***	**	*	*	****	***
5	*	*	*	*	****	****
6	**	**	**	***	*****	*****
7	**	**	*	**	****	****
8	*	*	*	**	*****	**
9	*	*	*	**	***	****
10	*	*	***	***	**	**
11	*	*	*	**	***	***
12	***	*	**	***	*****	*****

Table 31 represents the evaluation of the participants on whether those three reviews are original or fake. We can see that all participants thought that review 1 was fake except subjects 4 and 12. They provide three stars out of five, which concludes it is somewhat original. However, they changed their decision after having the explanations bu putting two and one star, respectively. For review 2, except for participants 1 and 2, everyone considered the review to be fake. Interestingly, review 3 were considered solely as original by the majority of the participants. However, after considering the explanations generated by the LRP-enabled explainability technique, two participants (subject 4 and 8) changed their decision by decreasing the mark.

Discussion on participants' opinion: We had a detailed discussion with each participant on how they came up with the decision whether a particular review was fake or original. For example, we asked the subject about the review 3. He said the following:

Subject 2: "The third review, because it was for me it was the most realistic. There was the name inside. So he seems to know the guy who's doing it and it's pretty, and it's really short."

He thought it was short and he believed in the text because it has a name. However, we asked him, what matters in predicting the review whether it is real or fake. He replied, its more about *linguistic form* (i.e., meaning grammatical structure and tone), not individual word.

Subject 2: "Yeah, and the second it's, uh, more about the words. And in the first, first, it's more about the linguistic form to me."

Similarly, Subject 3 also thought that highlighting relevant words as an explanation might not make sense in explaining review identification whether it is fake or real. It's about the whole text. He added, for the second review, based on the repetitive texts he identified review 2 as fake.

Subject 3: Uh, the second one is, I also think it's completely made up by AI, um, because it's very repetitive and, uh, uh, some sentences you just read and you think no human would write like this. Um, and then the third one to me also was the most realistic one because it is kind of short. It's very, it kind of seems authentic in terms of like the excitement.

Subject 3: "No, to me, it's not the singular words. To me, it really is the structure and the whole like, the thing as a whole that, um, makes it seem like it's AI generated"

Subjects 4 and 5 provided their insight about whether our generated explanations make sense to understand the decision. They both thought that the current form of explanation might help to some degree to comprehend the decision, 2 out of 10.

Subject 4: : "I think it might, it might make sense to some degree. But as, uh, my colleague just said, it's more about the, the overall. Two out of 10"

Subject 5: "No, no. Not from my part. I have to reread that, like out of 10. Uh, like two or three."

Subject 6 has found something very interesting in review 2, for example, information like age, and jobs are not relevant and these are not commonly used in review. He also identified that this is a very long review,

generally, people are too lazy to write.

Subject 6: "Because no matter, um, his age or his, um, job and something like this or for buying gloves, um, and also it's, um, too long. And I guess, um, people are, most people are lazy to write this kind of message. Yeah."

Grammatical information is identified as important to understand whether the review is fake or real by subject 7. He considered the more the number of adjectives that exist in the review, the more realistic the review is.

Subject 7: "No, just any adjective. So for example, the ones that I have rated the, the, the realest, have more objectives than, than the other ones. So it could be just, um, your personal opinion, it's not about."

He also thought the individual word might have some importance towards certain classes, but it should be the whole context of the review.

Subject 7: "For me, for me, they didn't really help me to find out if they are or not real. Uh, I think it's more like a context thing. Only, I mean for me the word has, has to, um, it's okay. It was the, the same."

Interestingly Subject 8 found our generated explanations are effective. Before accessing the explanations, subject 8 provided review 3 as 5 starts, meaning a fully real review. But after he went through the generated explanations, he thought this was also a fake review. Though it has several good adjectives, but he thinks these are the reasons to be fake, contradictory to subject 7.

Subject 8: "So, the third one actually was from me, at the first, um, I gave them five stars. So that it's likely to happen because it's short. I think in real life everyone just give short recommendations and not long recommendations. But then after reading the AI, um, explanation, yes. Um, reading the words excellent, amazing, great. I also think that it's, it's not a real review."

In summary, the explanations generated by LRP technique to highlight-

ing important relevant words in context of fake review identification can make general users sense in minimum scale. On contrary, for some application areas, for example, sentiment classification [19] and hate speech recognition [149], where LRP-based explanations are quite good to understand the reason behind the prediction. One of the main reasons identified through the empirical user study why explanations highlighting relevant words is the use of similar words in both fake and original review. For both positive and negative reviews, we observed similar adjectives or other praising or criticizing words are used in both fake and original reviews. For example, in sentiment analysis task, there are some terms or negation elements that are used for specific positive and negative class [19]. For another example, patent classification task [285], terms related to specific scientific fields are used in the patent text. Rather, in the context of fake review identification task, grammatical structure of the sentence, tone and overall context matters most in explaining the decisions.

12.5 Conclusion and Future Direction

In this paper, we proposed transparent and interpretable fake review detection framework applying transformer models including DistilBERT and XLNet. We also apply multiple deep learning models including BiLSTM, CNN, CNN-LSTM, and CNN-BiLSTM for modeling the fake review detection task. Then, we adopted LRP technique to open the the black-box deep learning model. The LRP can explain why a particular prediction has been made. We conducted experiments in multiple settings and applied our models to two different benchmark datasets. Based on the experimental results, we demonstrated that our proposed DistilBERT- and XLNet-based fake review detection models significantly outperformed other ensemble ML and DL models. Compared to the previously known related methods, our method also outperformed NBSVM,

OpenAI, and fakeRoBERTa methods on the same dataset. In the end, we demonstrated explanations provided by our adopted LRP technique for multiple example reviews for different categories. The empirical user evaluation with human subjects indicates further requirements to generate and present the explanations for any specific decision.

In the future, we are planning to have an empirical study to measure the quality of the generated explanations. Further, it would be interesting to consider the elicited requirements and findings from user evaluation for explanation generation.

Acknowledgment

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 955422.

Part V

Research Outcome and Conclusion

13 Discussion

This chapter presents the overall findings of the thesis and discusses how much we addressed the research questions we mentioned section 1.5 by mapping the findings and contributions. We present the observations on the experimental results of our proposed methods for three different application scenarios. Our research questions were about how the explanations vary across applications, what essential facts needed to consider in presenting explanations, and how we can achieve actionable explanations.

Overall, our research aimed to explore a wide range of application contexts and address the challenges of explaining AI models' decisions. We conducted extensive investigations into explainable AI methods across three different application domains and various sub-tasks, leading to a diverse array of findings and scenarios. In most application scenarios we explored, certain types of explanations may be meaningful, while in others, they may not be useful enough. On the other hand, explanations representation and their underlying information can also vary on the application contexts and user-expertise variability. In the following, we first discuss the observation and findings on three different application contexts and then we revisit the research questions with a broad discussion on how much we achieved.

13.1 Overall Findings

The introduced XAI methods applied to three different application scenarios achieved effectiveness in accurate predictions and can explain the predictions with a straightforward representation of the underlying facts and rationale. We rely on model-agnostic explainability techniques to explain the prediction; in almost all experiments, we first develop ML or DL predictive or forecasting models, and then XAI techniques come into play

to explain the models and their predictions [279, 280, 277, 285]. Across all applications, our introduced techniques achieved high performance in terms of standard evaluation metrics and compared to the existing approaches, the performances are significantly higher. This section will highlight the performance of proposed methods in two directions: (i) predictive performance and (ii) the evaluation and the observation of generated explanations.

13.1.1 Smart Home Applications

Overall smart home system is a combination of multiple AI-driven applications including energy demand forecasting, indoor temperature control systems and HVAC system [140, 76, 197] (chapter 5, 6, & 7). Energy demand forecasting for households with high precision needs extensive analysis of different features, consumption patterns for appliances, activities, and seasonal effects [164, 197]. DL-based multivariate forecasting model presented in chapter 6 can also predict the appliance level future consumption effectively as compared to the previous studies [161, 164], which can make household members aware of being more optimized in energy uses. On top of that, the explanations generated by our introduced DeepLIFT-enabled explainable forecasting framework can map the contribution of appliances and household activities with the corresponding time. The generated explanations can provide more fine-grained information about how and on which appliances their total cost associated with energy consumption is distributed.

The need for easily understandable explanations in the context of smart home systems analyzing two different application scenarios has been presented in chapter 5. The demonstration of explanations for different smart home applications highlighted the research gaps in why smart home users might need more fine-grained explanations. Considering the challenges and research gaps, chapter 5 sought three concepts of

syntax-, semantic- and pragmatic-level explanations. To achieve such, HCI techniques, including user studies, prototyping, technology probes analysis, and heuristic evaluation, are needed to apply for generating better understandable explanations for users.

The explanations in terms of household activities responsible for future energy consumption might solve the user experience variability challenges by providing more accessible explanations than highlighting the contributions of every feature. Explanations using household activities can be more beneficial in making sense of what household activities might be responsible to future energy consumption. Considering the presented data in explanations, they might change and optimize their energy consumption practice.

The quality of the generated explanations can be evaluated based on different criteria, including the contents, representation, and user satisfaction [219]. There are several properties including correctness [15], completeness [73], and consistency [210] that need to be considered to evaluate the content of the explanations. On the other end, compactness, composition, and confidence are other important properties to judge the quality of the explanations representation [41, 219].

In chapter 6, the performance of the introduced household energy demand forecasting system is effective and can predict the appliance level consumption with better accuracy compared to the previous work [161, 163]. The generated explanations were evaluated by introducing a new evaluation metric. Previously used metrics are often used for classification tasks. However, the application context and features in energy demand forecasting are pretty different. After carefully analyzing the features, activities, and seasonality with the associated time frames, we needed to evaluate the efficiency of the explanations employing a correlation coefficient. The correlation coefficient can identify whether the identified contributions of different features or activities in explanations

align with the ground truth.

The explanations generated were evaluated by introducing a new metric, *contribution monotonicity coefficient, CMC*, that can compute whether the highlighted contributions are aligned with past consumption patterns. This metric can evaluate how accurate the facts used in the representation of the explanations [219]. Compared with the ground truth, this evaluation metric can measure the efficiency of the explanations in the context of time series forecasting. The explanations represented with relevant household activities indicated the efficiency of the generated explanations and aligned with their findings compared to the existing approach to household activity recognition [302]. Moreover, the experiments on two different datasets concluded that the introduced metric can be used to evaluate the effectiveness of the explanations of other multivariate time series forecasting tasks.

One of the previous studies applied methods to explain the prediction from energy demand forecasting systems with heatmap-based explanations [163]. However, the generated explanations are highly technical and could help practitioners design new models and improve performance. The comparison of the overall findings of explanations generated using household activities is also analogous to the findings of the household activity recognition from the energy consumption by [302].

Generally, modeling personal thermal comfort preference for an HVAC system depends on high dimensional features, which might be costly to implement since it requires various physiological data from the occupants in the home [320, 191]. This requirement leads to a higher cost of installing expensive sensors. The predictive models are also computationally expensive when the data samples have much more features to consider. However, the findings for personal thermal comfort preference prediction system in chapter 7 are pretty interesting. The preliminary analysis and exploratory experiments concluded that several redundant

and irrelevant features in the dataset might hinder prediction performance. Following the preliminary findings and hypothesis, the use of supervised feature selection techniques was effective in discarding noisy and irrelevant features.

The data inadequacy was another challenge in modeling the thermal comfort preference prediction task. The datasets often need to be more prominent in size, more in terms of several samples to train ML models for personal preference prediction. On the other hand, collecting data from human subjects is expensive and lengthy. To overcome this challenge, conditional tabular generative adversarial networks (CTGAN) were introduced to create synthetic samples considering the existing samples. With this, the dataset had enough samples to train the model. The experimental results in multiple directions demonstrated the effectiveness of the introduced feature selection and GAN compared to prior works [78, 334].

The feature selection techniques were effective in presenting the explanations because they discard irrelevant and redundant features. Hence, the explanations only consider features that are highly relevant to the decisions. In this way, feature selection can be used in presenting explanations, and hence, explanations can be more focused. Similar findings have also been reported in other applications for generating explanations for high-dimensional datasets [151].

13.1.2 Business Applications

The objectives of having explainable business applications might be related to both the owner or stakeholders of the business and the end users or clients of the applications (chapter 9). Therefore, the explanations might play different roles for the different users [251]. In the area of e-commerce, we demonstrated the explanations for stakehold-

ers of the retail company on how they overcome future product back-orders by considering the prediction and the explanations with extracted facts. Eventually, the application's performance should be accurate, besides providing explanations. This dissertation also contributed to introducing a high-performance, explainable product back-order prediction system. The CNN-based explainable model for product back-order prediction improved significantly compared to existing methods. Since the back-order is a rare event in inventory management, the dataset is extremely imbalanced. The imbalance in the dataset leads to model bias and might hinder the high predictive performance. Our method of applying Adaptive Synthetic oversampling (ADASYN) [185, 123] to get rid of the imbalance problem performs better than the other existing techniques, including SMOTE [62, 119, 273]. The overall performance has been evaluated based on AUC, the metric that can compute the performance of the predictive models on a dataset with imbalances. Working with such an imbalanced dataset requires decent preprocessing steps, including handling missing values and normalization, before applying the deep learning models.

Our generated global and local explanations, applying SHAP and LIME, can make sense of the model's overall priorities and the reason behind specific predictions. Moreover, it can pinpoint the features responsible for the future back-order of specific products. From the stakeholders' point of view, the explanations can be further analyzed to know what are the most responsible features that can be changed to overturn the decision; hence, in the future, that particular product will not be back-ordered, and it will ensure the revenue and decrease the possible loss.

13.1.3 NLP Applications

The explanations for NLP tasks are generally generated considering the relevant terms towards the prediction model and classification mod-

els [211]. In this dissertation, two different types of explainable text classification methods have been introduced, one with DL models using the semantic representation from word embedding and another one based on transformers for two different NLP applications, patent classification and fake review identification, respectively (chapter 11 & 12). DL models, including BiLSTM and CNN-BiLSTM, performed well in classifying scientific patents and outperformed previous existing methods on two different patent classification datasets [252, 183]. On the other hand, transformer-based fake review identification models could also identify the review types effectively with high precision, compared to existing models [259].

However, the effectiveness of the explanations generated by our introduced LRP techniques differs for the context of the two applications. In the case of scientific patent classification, we observed that the explanations help understand why a particular patent is classified into a class. The explanations can map the related scientific terms related to the class. For example, the explanations identified the words related to *chemistry* patents, which are not generally used in other patents. Along with identifying the relevant scientific terms for the corresponding patents, it can emphasize the degree of relevancy of particular terms to the class in different visualizations.

The decisions have been explained by highlighting the corresponding related terms for the fake review identification task. From the preliminary analysis of the generated explanations for this application context, we have observed that the identified terms for fake and original reviews are similar. This is because both fake and original reviews generally use similar words, either to praise or criticize a particular product. Therefore, explanations highlighting only terms with the degree of relevancy might not be efficient in this particular application.

To gain more insight into the generated explanations and determine

whether they are helpful to understanding the AI decision or not, our conducted user study on evaluating the explanations with human subjects suggested that the explanations can make sense on a small scale. The findings of the empirical user study concluded that the explanations behind any particular reviews, be they fake or original, are not related to only the terms used. Instead, it depends on the overall tone of the sentences or text of the review and the grammatical structures.

13.2 Revisiting Research Questions

13.2.1 Achieving high-performance

Before exploring explainability in different application scenarios, first, we needed to achieve high-performance ML models that provide accurate prediction in every application. Here, we discuss how we tried to answer the first research question and what strategies and techniques were applied to overcome the challenges, such as data inadequacy and data imbalances in achieving high-performance ML models.

- **RQ 1:** *What techniques and strategies can be employed to achieve high-performance ML models addressing technical challenges such as data imbalance, data inadequacy, and model bias?*

We explored and proposed high-performance ML models for smart home systems for two major applications, including energy demand forecasting and personal thermal comfort preference prediction tasks. From these two applications, thermal comfort preference prediction datasets need more data; there was a shortage of adequate data needed to train ML model [78] effectively. Collecting and annotating high-dimensional data from human subjects is also time-consuming and costly. We analyzed the available data carefully and introduced a conditional tabular GAN architecture to generate synthetic data to achieve data adequacy

for model training (chapter 7). The previously studied method also suffers from data inadequacy challenges [191]. Since the data were high-dimensional, the careful investigation applying a correlation coefficient indicated that the data had too many high-correlated and redundant features. Therefore, we introduced multiple supervised feature selection techniques to filter such features. With synthetic data generation using CTGAN and feature selection techniques, we achieved high-performance personal thermal comfort preference prediction models compared to previous known related methods [191, 78, 25].

When it comes to forecasting energy demand for smart homes, we took a different approach. Recognizing that the forecasting problem often depends on seasonal effects, we extracted seasonality features. These features, when combined with the energy consumption from different appliances, were used to train the widely known LSTM model for weekly and monthly prediction of the overall energy consumption. The results were promising. We achieved better performance in two different datasets consisting of five different household energy consumption data [164, 302]. The generated explanations not only validated our approach but also indicated that the introduced features were contributing significantly to the overall prediction.

In the business problem of product backorder prediction, one of the significant challenges was that the dataset was extremely imbalanced. Therefore, the model is expected to be biased on the majority class samples. We applied several approaches to address the extreme class imbalance problem and found that ADASYN oversampling would be the best one to tackle data imbalance, particularly for modeling product backorder prediction [286, 119]. Eventually, our proposed CNN-based product backorder prediction model employing the data generated by the ADASYN oversampling technique achieved a new state-of-the-art performance and outperformed known related methods [286, 119, 226, 134].

Applying DL and transformer-based classification models, we modeled two different NLP applications before explaining the predictions. Compared to the previous methods on both tasks, we achieved higher performance due to the data imbalance, and our methods outperformed other related methods that applied DL and transformer-based models to classify scientific patents and detect fake reviews in e-commerce applications [252, 183, 259].

13.2.2 Explanations vary across applications' context

The next two research questions was about how we can achieve human-centered explainability for given different applications scenario and contexts. We mostly conducted experiments applying different XAI techniques in the application areas including smart homes, e-commerce and NLP.

- **RQ 2**: *How can human-understandable explainability be achieved for ML models within a specific application domain?*
- **RQ 3**: *How do explanations vary across different real-world applications?*

For smart home application scenarios (chapter 5, 6 & 7), namely in energy demand forecasting and personal thermal comfort preference prediction, the outcome of our introduced methods and explainability techniques concluded that even in the same application area, the explanations types and representations might be different. For energy demand forecasting, explanations for general users highlighting energy consumption by different appliances and time of use might make sense as to how their overall energy bills are associated with using different appliances. They can find the distribution and pattern of energy uses associated with the appliances.

To make more holistic explanations in more accessible forms, we empirically observed that the explanations highlighting the contributions of different household activities are adequate to understand. It provided an overall view of how the household members consume energy in different activities such as cooking and watching TV. However, considering the cost associated with the energy consumption of different appliances and time of use might make another dimension, and inhabitants then can optimize their energy consumption practices. It was also evident that the associated cost might matter most on the energy consumption in households in different activities (chapter 6).

The explanations help general smart home users understand why they need a specific amount of energy for the upcoming month/week. The explanations in terms of activities are also usable to change the consumption practice since these might provide clear indications about the responsible activities on total consumption. Therefore, users might consider optimizing a specific activity (i.e., the most consumed one) by changing their consumption routine. Compared to the previous studies on household energy demand forecasting, our findings contributed to explaining decisions in easily understandable representations. In contrast, previous methods [57, 162, 305] barely investigated explainability except in one that provided technical explanations [163].

However, on explainable product backorder prediction tasks, the users are not lay people but the stakeholders and inventory managers of the retail company [119, 226]. So, the explanations only highlight why a particular product will go back-ordered. However, they might expect to know the factors and how they can overturn the prediction considering the explanations.

In the explainable patent classification problem, explanations in terms of heatmaps and word clouds are effective in understanding the reason behind the prediction of the patent class. This is because the LRP-

enabled explainability technique is capable of identifying relevant, important scientific terms corresponding to the particular patent.

On a similar application, fake review identification and explanations with similar LRP-enabled explainability techniques could be more helpful in comprehending why a particular e-commerce review is predicted as fake. Though it can identify the relevant important terms and present them in terms of heatmap and word cloud, the user evaluation of such explanations concluded that fake reviews depend on the structure and tone of the sentences, not the terms. This is an interesting finding that goes opposite to the observation on patent classification.

In summary, the investigation of demonstrating explanations with multiple application areas introducing several explainability techniques concluded that we must study the users' and stakeholders' needs. After careful analysis and consideration, we must select and develop explainable systems. Interestingly, for the same text classification application, the explanations in patent classifications with the same XAI techniques help comprehend the prediction. However, explanations generated using same techniques for fake review identification only makes a little more sense.

13.2.3 Generating Explanations considering Underlying Facts

Generating explanations considering the facts extracted by the XAI techniques for the overall global model's priorities and specific predictions depends on the application and users' or stakeholders' context. Our proposed ML models and explainability techniques in different application scenarios also concluded with similar findings.

- **RQ 4:** *What underlying facts and rationale should we consider when generating explanations within application scenarios?*

For the energy demand forecasting system (chapter 5 & 6), DeepLIFT-

based explainability techniques consider both the previous time frames and the appliances' consumption. Because the deep multi-variate forecasting model depends on the time sequence when predicting the total consumption for the upcoming week/month. Therefore, we must consider the time and consumption of appliances (features) in generating the explanations. However, explanations of the intensity of energy consumption by different activities in households can make sense to understand more easily, depicting how their total consumption is distributed among different activities. Therefore, for explaining prediction for energy demand forecasting applications in smart homes, the relevant underlying facts that could be easier for users to understand include the consumption of appliances and activities, the time when the users might use to consume more energy [113].

For global interpretability for the personal thermal comfort preference prediction task, all features, both environmental and physiological, should not be considered. Many features are not controllable, such as weather information (i.e., outside temperature). Though explanations using uncontrollable features can not provide insight to take necessary actions, they can help users make sense.

The explanations for the predictions from our proposed CNN-based backorder prediction model (chapter 9) and the overall global explanations can identify the most critical relevant features for the overall model's decision-making. However, for local explanations for specific predictions, we observed that some features are not controllable, meaning the stakeholders can not take the necessary steps to overturn the prediction to minimize future loss. The generated explanations are helpful to some extent, at least in knowing the reason, but those not changeable features might be omitted for explanation generation. For example, the shipping time can be an essential feature in inventory management [119, 286], the explanations with the hint to increase or decrease

the product shipping time would be somewhat more helpful for the inventory manager to take necessary steps.

Since the priority from the stakeholder side is to increase the revenue and decrease the company's loss due to the possible backorder situation of a product [119], the features associated with cost should be given importance in generating explanations. In particular, explanations are expected to make the decision more transparent to the stakeholders so they can have clear hints on overturning it and making it financially profitable. For example, explanations can clarify what possible financial changes will happen if stakeholders change some policy, such as adjusting lead time. At the same time, stakeholders need to be aware of customer satisfaction. Therefore, explanations highlighting changing features, such as extending the shipping time to overturn the backorder situation, would not be ideal for the users.

On NLP applications, in patent classification and fake review identification (chapter 11), the facts to consider for generating explanations are contrary to our observation. The explanations generated for specific predictions of a patent are helpful when they can identify the relevant critical scientific terms. For explainable patent classification, the generated explanations applying word cloud and heatmap can illustrate why a particular patent is classified to a specific class. The highlighted terms used in the explanations are also helpful when users know about the degree of relevance of the prediction. Despite a few exceptions, the relevant identified terms towards the class and their degree of relevancy can explain the underlying reasons why a patent is associated with the particular patent.

However, in the same way, and applying the same LRP-enabled explainability technique, the generated explanations for fake review identification also represented highlighted weighted terms (chapter 12). However, in fake review identification contexts, similar explanations are incapable

of making sense compared to the patent classification task. Because every patent has different words associated with patent classes. However, in the fake review identification dataset, the reviews written by AI tools also have terms often used by human subjects. That is why the reviews written by human users and AI tools might have similar words to praise or criticize certain products. In turn, the prediction, whether the review is fake or original, might not be related to the words but to the sentence's structure and the language's tone. The empirical study with human subjects also concluded that in the case of fake review identification, the identified words do not play a vital role in being classified as fake or original but the overall structure or tone of the reviews.

13.2.4 Achieving Actionable Explanations

In the ideal case, the explanations should not only be self-explanatory to understand the models and their predictions but also it should be actionable [200]. By actionable, explanations should understand the decision and provide insight to take corrective actions to the users. Technically, explanations often provide insights and facts that can be useful for the AI practitioner to modify the ML models to improve the model's performance. For example, if we consider the explanations for deep energy demand forecasting systems, inhabitants should be able to use the findings to optimize their energy consumption routine and change their energy activity. This dissertation also focused on how we could generate actionable explanations through exploratory explanations.

- **RQ 5:** *How can actionable explanations be generated to optimize practice for given application contexts?*

In the case of NLP applications (chapter 11 & 12), from the discussion in the previous section13.2.3, we presented that the prominent widely used LRP-enabled explainability techniques provide meaningful explanations

for patent classification. Eventually, explanations can be utilized to determine why a patent has been classified into a specific class. Moreover, the degree of relevant terms has also been identified as terms with the highest influence on a specific class. Experts who work on classifying patents and scientists in different scientific areas working on writing patents might have substantial ideas about the reason for the classification of patents. In turn, the scientists can consider the generated explanations to leverage their preferred terms in their patents. Besides, the professionals who work for patent classification could have used the explanations and be satisfied to assign a particular class to a patent.

However, for the fake review identification task (chapter 12), the same LRP-enabled explainability techniques suffer to make understand why the specific reviews are fake or original. In this case, the successful explainability techniques are not working correctly to make sense of the prediction. The challenges we mentioned in section 1.4 are evident. One specific XAI technique is not enough even for similar tasks, and the application context played a vital role in understanding the facts. Through the empirical evaluation with human subjects, it can be concluded that the HCI techniques, such as user study and prototyping discussed in chapter 5 should be considered before applying explainability techniques. Otherwise, it might provide explanations that can be useful to know the models' priority but will not be actionable. The empirical evaluation reveals that the grammatical structure and sentence tone should be used to generate actionable explanations.

Since the explanations in energy demand forecasting systems are related to two different dimensions, features and time (chapter 5 & 6), there are some features that we tried to extract depending on the seasonality. However, the explanations generated by DeepLIFT can depict the distributions of energy consumption in different household activities. The explanations can highlight the contribution of different appliances and

household activities responsible for energy consumption. Total household energy consumption, distributed by the intensity of different household activities, is more understandable than consumption distributed across the appliances [302]. The explanations regarding the intensity of activities are more actionable than the appliance-level consumption-based explanations. With such explanations, household users might take corrective action to optimize their total energy consumption. It would be even further beneficial to the household inhabitants if explanations could provide information highlighting some actions based on their historical data that would help optimize energy consumption. Still, the current form of explanations can be improved by incorporating the users' feedback by providing an interactive, collaborative interface. For actionable explanations in thermal comfort preference prediction, occupants might expect information so that they can change particular attributes not related to energy consumption. Hence, it reduces the associated energy cost and improves the comfort inside the home.

For explainable business systems (chapter 9), actionable explanations are required to provide insight clearly that the stakeholder of the company can make policy-level decisions so that both the revenue of the company and the satisfaction level of the customer would be higher. Our proposed explainable product backorder prediction system can highlight essential features related to a particular product. Suppose a particular product is predicted as a backordered one. In that case, the explanations associated with this prediction can depict the relevant feature to the inventory manager, which can be analyzed for future changes [119, 286]. However, some features and attributes can not be changed immediately; those are called uncontrollable, but the manager can think of changing the controllable feature to see how that feature contributed to the prediction. For example, the inventory amount and shipping time can be features that have a higher influence on the prediction. So, changing

such feature values for specific products back and forth might provide critical insight that can be implemented to reverse future backorders.

14 Conclusion

14.1 Summary of the Dissertation

This dissertation focused on generating easy-to-understand actionable explanations for general users so that they can understand the decisions made by complex AI models and take corrective actions accordingly. Therefore, different application scenarios have been investigated to explore what facts and rationale should be considered in generating explanations, how explanations vary across different applications, and how explainability could be actionable. Keeping these as the main focus, this dissertation has five parts.

The first part (Part I) is about the introduction & and overview of the thesis, presenting the motivations and research challenges underlying achieving explainable AI-enabled systems. It then presents the goal of this thesis by mentioning the research questions that need to be addressed. The second chapter (chapter 2) of the first part discussed related works on the overall progress of technical and human-centered XAI. It highlighted the state-of-the-art research on three selected application areas: smart home, business, and NLP. The last chapter (chapter 3) of part I presented the selection of application areas, studied research methodology, and outline of the study.

Next, part II presents the proposed approaches to make smart home sub-tasks explainable. Chapter 5 advocated the necessity of applying explainable methods in smart home application areas, demonstrating the experiments, and analyzing scenarios. It further elicited the research challenges and provided high-level research directions to consider HCI methodology to make explanations user-centric. In the next chapter (chapter 6), this thesis presented a novel XAI technique for deep multi-variate time series, an energy demand forecasting model, and an evaluation metric to compute the efficiency of the generated explana-

tions compared to ground truth. The last chapter (chapter 7) of this part was about modeling thermal comfort preference, which introduced feature selection techniques to discard irrelevant features, GAN to generate data samples to overcome the data inadequacy challenge synthetically, and finally, the global interpretability of thermal comfort preference prediction models.

Part III focused on interpreting the models in the e-commerce area. This part presented a new explainable CNN-based product backorder prediction method that can efficiently predict future backorder by addressing the extreme data imbalance challenge by introducing the ADASYN oversampling technique. To explain the prediction from the model, it can present the explanations for both global and local contexts. Applying two widely used model-agnostic XAI techniques, SHAP and LIME, it presented the explanations in different forms and purposes.

For explainable models in NLP, part IV presented interpretable models for explaining the decisions for fake review detection (chapter 12) and patent classification tasks (chapter 11). For both cases, the introduced LRP explainability techniques can explain the prediction by identifying relevant important terms with their degree of relevancy associated with the predicted decisions. With an empirical user study, the efficiency of the explanations for the fake review detection system has been evaluated, and it was found that the explanation in this context might related to the tone and grammatical structure, not related to the related words.

In the final part, this dissertation discussed the findings and contributions by revising the research questions. It then concludes the thesis with potential limitations and possible future works.

14.2 Contributions

The broad contribution of this dissertation is related to the objective to achieve easy-to-understand explanations for AI-enabled systems investigation by answering the research questions mentioned in 1.5. We explored three application areas to investigate how explanations vary across applications, how the user expertise variability matters for generating explanations, and what things should be considered to achieve actionable and easy-to-understand explanations. The contributions of this dissertation are summarized in three application areas.

In **smart home application** domain, we investigated and sought the needs of human-centered explanations by demonstrating the experimental results with state-of-the-art approaches. We elicited challenges towards achieving human-centered explanations and the research directions considering HCI techniques that can be studied and applied in other applications (chapter 5). For a household energy demand forecasting system, this thesis proposed an explainable deep multi-variate time series forecasting technique for an energy demand forecasting system. We modified DeepLIFT to map the contribution of different features corresponding to the time. The introduced evaluation metric, *CMC*, can be applied to measure the effectiveness of explanations from other time series forecasting models (chapter 6).

The performance of integration of feature selection techniques in modeling personal thermal comfort provides a broad direction for developing any predictive systems with high dimensional data. Our introduced CTGAN can be applicable to overcome the data inadequacy challenge for any other applications, which can be a pre-step before training robust ML models (chapter 7). Moreover, the explanations from ML models that incorporate data inadequacy and problems with high dimensional features can be practical in two ways. The explanations might not be biased since it solves the data inadequacy and imbalance problem. On

the other hand, effective supervised feature selection techniques should help represent facts in terms of explanations. The explanations might be more comprehensive because feature selection techniques filter out irrelevant features.

In **business application** domain, this thesis proposed a new explainable CNN-based product backorder prediction method applying model-agnostic global and local explainability techniques. The model demonstrated that the explanations can capture the facts behind specific predictions and explain the model's overall decision-making priorities with global explanations (chapter 9). The study also introduced the idea that explanations should be more pinpointed so that the manager can take controllable necessary steps.

Finally, for **NLP applications**, this thesis demonstrated interesting findings that in the same text classification domain, the explanations could be practical for the patent classification system (chapter 11). In contrast, the explanations from the same XAI technique in fake review identification can be less productive (chapter 12). However, our proposed DL and transformer-based classification models performed significantly better in patent classification and fake review identification tasks, respectively. The conducted user study on evaluating explanations from fake review identification systems indicates that competency and application objectives should be considered to represent explanations.

14.3 Limitations & Future Work

The potential limitations of this dissertation will be discussed in the remainder of this chapter, along with possible future work that might be the next step in achieving human-centered explainability, ensuring that general users have a complete understanding of explanations. Therefore, this would enhance the real-world adoption of AI.

This dissertation employed an experimental approach to applying various explainability techniques in different applications to understand how we can achieve this with current progress in XAI. Therefore, the following points and areas might be interesting to investigate to fulfill the requirements and implement human-centered explainability.

As noted earlier, one explanation technique only fits some. Therefore, the adoption of multi-purpose explanation techniques is not just ideal, but necessary for specific real-world applications. These techniques offer explanations that consider the variability in stakeholders' experience and the information they need to comprehend particular decisions. Since smart home applications are more user-centric and there is a high user experience variability to comprehend AI decisions, employing multi-purpose explanation interfaces would empower users to choose the explanation types and understand the decisions better.

The introduction of Generative AI [304] and large language models (LLMs) [342], a recent breakthrough, holds great promise in fine-tuning the explanations. With their ability to understand and generate language based on a command, LLMs can be used to represent and generate explanations. An empirical study with end-users could uncover the potential of LLMs, offering a promising outlook. Furthermore, a knowledge graph [84, 176] can be an asset for fine-graining the generated explanations, further enhancing the potential benefits.

Another potential limitation of this dissertation is the pressing need for large-scale user studies in different problem domains to understand how the information needed for explanations varies across application areas and users. Conducting larger-scale user studies and applying those requirements to generate explanations would be an exciting research area.

We relied on technical experiments since we focused on generating user-understandable explanations by introducing XAI techniques and using facts and rationale to represent the explanations. This is another lim-

itation of our work, as we have yet to explore designing interactive interfaces [322, 321] to enhance user engagement. Therefore, this could be a potential area for future work in designing and developing more interactive interfaces to present the explanations.

References

[1] Mahmoud M Abdelrahman, Adrian Chong, and Clayton Miller. 2022. Personal thermal comfort models using digital twins: Preference prediction with BIM-extracted spatial–temporal proximity data from Build2Vec. *Building and Environment* 207 (2022), 108532.

[2] Ashraf Abdul, Christian von der Weth, Mohan Kankanhalli, and Brian Y Lim. 2020. COGAM: measuring and moderating cognitive load in machine learning model explanations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[3] Babak Abedin, Mathias Klier, Christian Meske, and Fethi Rabhi. 2022. Introduction to the Minitrack on Explainable Artificial Intelligence (XAI). In *Proceedings of the 55th Hawaii International Conference on System Sciences*. 1,2. `http://hdl.handle.net/10125/70765`

[4] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access* 6 (2018), 52138–52160.

[5] Amina Adadi and Mohammed Berrada. 2020. Explainable AI for healthcare: from black box to interpretable models. In *Embedded Systems and Artificial Intelligence: Proceedings of ESAI 2019, Fez, Morocco*. Springer, 327–337.

[6] Surabhi Adhikari, Surendrabikram Thapa, Usman Naseem, Hai Ya Lu, Gnana Bharathy, and Mukesh Prasad. 2023. Explainable hybrid word representations for sentiment analysis of financial news. *Neural Networks* 164 (2023), 115–123.

[7] David W. Aha, Dennis Kibler, and Marc K. Albert. 1991. Instance-Based Learning Algorithms. *Machine Learning* 6 (1991). Issue 1. https://doi.org/10.1023/A:1022689900470

[8] Jawad Ahmad, Ahsen Tahir, Hadi Larijani, Fawad Ahmed, Syed Aziz Shah, Adam James Hall, and William J Buchanan. 2020. Energy demand forecasting of buildings using random neural networks. *Journal of Intelligent & Fuzzy Systems* 38, 4 (2020), 4753–4765.

[9] Rajendra Akerkar. 2019. *Artificial intelligence for business*. Springer.

[10] Fadi Al-Turjman. 2019. *Artificial intelligence in IoT*. Springer.

[11] Alper T Alan, Mike Shann, Enrico Costanza, Sarvapali D Ramchurn, and Sven Seuken. 2016. It is too hot: An in-situ study of three designs for heating. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 5262–5273.

[12] Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. 2023. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information fusion* 99 (2023), 101805.

[13] Usman Ali, Mohammad Haris Shamsi, Cathal Hoare, Eleni Mangina, and James O'Donnell. 2021. Review of urban building energy modeling (UBEM) approaches, methods and tools using qualitative and quantitative analysis. *Energy and buildings* 246 (2021), 111073.

[14] Saleh Nagi Alsubari, Sachin N Deshmukh, Theyazn HH Aldhyani, Abdullah H Al Nefaie, and Melfi Alrasheedi. 2023. Rule-based clas-

sifiers for identifying fake reviews in e-commerce: a deep learning system. In *Fuzzy, Rough and Intuitionistic Fuzzy Set Approaches for Data Handling: Theory and Applications*. Springer, 257–276.

[15] David Alvarez Melis and Tommi Jaakkola. 2018. Towards robust interpretability with self-explaining neural networks. *Advances in neural information processing systems* 31 (2018).

[16] David Alvarez-Melis and Tommi S Jaakkola. 2017. A causal framework for explaining the predictions of black-box sequence-to-sequence models. *arXiv preprint arXiv:1707.01943* (2017). https://doi.org/10.48550/arXiv.1707.01943

[17] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein generative adversarial networks. In *International conference on machine learning*. PMLR, 214–223.

[18] Segun Taofeek Aroyehun, Jason Angel, Navonil Majumder, Alexander Gelbukh, and Amir Hussain. 2021. Leveraging label hierarchy using transfer and multi-task learning: A case study on patent classification. *Neurocomputing* 464 (2021), 421–431.

[19] Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2017. " What is relevant in a text document?": An interpretable machine learning approach. *PloS one* 12, 8 (2017), e0181142.

[20] Leila Arras, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2017. Explaining recurrent neural network predictions in sentiment analysis. *arXiv preprint arXiv:1706.07206* (2017).

[21] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020.

Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion* 58 (2020), 82–115.

[22] Luca Arrotta. 2021. Multi-inhabitant and explainable Activity Recognition in Smart Homes. In *2021 22nd IEEE International Conference on Mobile Data Management (MDM)*. IEEE, 264–266.

[23] Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilović, et al. 2019. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *arXiv preprint arXiv:1909.03012* (2019). https://doi.org/10.48550/arXiv.1909.03012

[24] Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilović, et al. 2021. AI Explainability 360 Toolkit. In *Proceedings of the 3rd ACM India Joint International Conference on Data Science & Management of Data (8th ACM IKDD CODS & 26th COMAD)*. 376–379.

[25] Ashrant Aryal and Burcin Becerik-Gerber. 2020. Thermal comfort modeling when personalized comfort systems are in use: Comparison of sensing and learning methods. *Building and Environment* 185 (2020), 107316.

[26] Ashrant Aryal, Burcin Becerik-Gerber, Gale M. Lucas, and Shawn C. Roll. 2021. Intelligent Agents to Improve Thermal Satisfaction by Controlling Personal Comfort Systems under Different Levels of Automation. *IEEE Internet of Things Journal* 8 (2021), 7069–7100. Issue 8. https://doi.org/10.1109/JIOT.2020.3038378

[27] Roy Assaf and Anika Schumann. 2019. Explainable deep neural

networks for multivariate time series predictions.. In *IJCAI*. 6488–6490.

[28] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one* 10, 7 (2015), e0130140.

[29] Snehasish Banerjee and Alton YK Chua. 2023. Understanding online fake review production strategies. *Journal of Business Research* 156 (2023), 113534.

[30] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[31] Rodrigo Barbado, Oscar Araque, and Carlos A Iglesias. 2019. A framework for fake review detection in online consumer electronics retailers. *Information Processing & Management* 56, 4 (2019), 1234–1244.

[32] Alejandro Barredo Arrieta, Sergio Gil-Lopez, Ibai Laña, Miren Nekane Bilbao, and Javier Del Ser. 2022. On the post-hoc explainability of deep echo state networks for time series forecasting, image and video classification. *Neural Computing and Applications* 34, 13 (2022), 10257–10277.

[33] Ivana Bartoletti. 2019. AI in healthcare: Ethical and privacy challenges. In *Conference on Artificial Intelligence in Medicine in Europe*. Springer, 7–10. https://doi.org/10.1007/978-3-030-21642-9_2

[34] Gourav Bathla, Pardeep Singh, Rahul Kumar Singh, Erik Cambria, and Rajeev Tiwari. 2022. Intelligent fake reviews detection based on aspect extraction and analysis using deep learning. *Neural Computing and Applications* 34, 22 (2022), 20213–20229.

[35] Ransome Epie Bawack, Samuel Fosso Wamba, Kevin Daniel André Carillo, and Shahriar Akter. 2022. Artificial intelligence in E-Commerce: a bibliometric study and literature review. *Electronic Markets* (2022), 1–42. https://doi.org/10.1007/s12525-022-00537-z

[36] Joseph Beck, Mia Stern, and Erik Haugsjaa. 1996. Applications of AI in Education. *XRDS: Crossroads, The ACM Magazine for Students* 3, 1 (1996), 11–15.

[37] Hamid Bekamiri, Daniel S Hain, and Roman Jurowetzki. 2021. PatentSBERTa: a deep NLP based hybrid model for patent distance and classification using augmented SBERT. *arXiv preprint arXiv:2103.11933* (2021).

[38] Andrew Bell, Ian Solano-Kamaiko, Oded Nov, and Julia Stoyanovich. 2022. It's just not that simple: an empirical study of the accuracy-explainability trade-off in machine learning for public policy. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 248–266.

[39] Claudio Bettini, Gabriele Civitarese, and Michele Fiori. 2021. Explainable Activity Recognition over Interpretable Models. In *2021 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*. IEEE, 32–37.

[40] Aakash Bhandary, Vruti Dobariya, Gokul Yenduri, Rutvij H Jhaveri, Saikat Gochhait, and Francesco Benedetto. 2024. En-

hancing Household Energy Consumption Predictions through Explainable AI Frameworks. *IEEE Access* (2024).

[41] Umang Bhatt, Adrian Weller, and José MF Moura. 2020. Evaluating and aggregating feature-based model explanations. *arXiv preprint arXiv:2005.00631* (2020).

[42] Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. 2016. Layer-wise relevance propagation for neural networks with local renormalization layers. In *Artificial Neural Networks and Machine Learning–ICANN 2016: 25th International Conference on Artificial Neural Networks, Barcelona, Spain, September 6-9, 2016, Proceedings, Part II 25*. Springer, 63–71.

[43] Şule Öztürk Birim, Ipek Kazancoglu, Sachin Kumar Mangla, Aysun Kahraman, Satish Kumar, and Yigit Kazancoglu. 2022. Detecting fake reviews through topic modelling. *Journal of Business Research* 149 (2022), 884–900.

[44] Jerzy Błaszczyński and Jerzy Stefanowski. 2015. Neighbourhood sampling in bagging for imbalanced data. *Neurocomputing* 150 (2015), 529–542. https://doi.org/10.1016/j.neucom.2014.07.064

[45] Stefan Blücher, Johanna Vielhaben, and Nils Strodthoff. 2022. PredDiff: Explanations and interactions from conditional expectations. *Artificial Intelligence* 312 (2022), 103774.

[46] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics* 5 (2017), 135–146.

[47] Judy Borowski, Roland S Zimmermann, Judith Schepers, Robert Geirhos, Thomas SA Wallis, Matthias Bethge, and Wieland Bren-

del. 2020. Exemplary natural images explain CNN activations better than state-of-the-art feature visualization. *arXiv preprint arXiv:2010.12606* (2020).

[48] Clara Bove, Jonathan Aigrain, Marie-Jeanne Lesot, Charles Tijus, and Marcin Detyniecki. 2022. Contextualization and exploration of local feature importance explanations to improve understanding and satisfaction of non-expert users. In *27th international conference on intelligent user interfaces*. 807–819.

[49] Gail Brager, Hui Zhang, and Edward Arens. 2015. Evolving opportunities for providing thermal comfort. *Building Research & Information* 43, 3 (2015), 274–287.

[50] Zana Buçinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th international conference on intelligent user interfaces*. 454–464.

[51] Niklas Bussmann, Paolo Giudici, Dimitri Marinelli, and Jochen Papenbrock. 2021. Explainable machine learning in credit risk management. *Computational Economics* 57, 1 (2021), 203–216. https://doi.org/10.1007/s10614-020-10042-0

[52] Yue Cao, Thomas Andrew Geddes, Jean Yee Hwa Yang, and Pengyi Yang. 2020. Ensemble deep learning in bioinformatics. *Nature Machine Intelligence* 2, 9 (2020), 500–508. https://doi.org/10.1038/s42256-020-0217-y

[53] Fabrizio Carcillo, Yann-Aël Le Borgne, Olivier Caelen, Yacine Kessaci, Frédéric Oblé, and Gianluca Bontempi. 2021. Combining unsupervised and supervised learning in credit card fraud detection. *Information sciences* 557 (2021), 317–331. https://doi.org/10.1016/j.ins.2019.05.042

[54] Emerson F Cardoso, Renato M Silva, and Tiago A Almeida. 2018. Towards automatic filtering of fake reviews. *Neurocomputing* 309 (2018), 106–116.

[55] Nico Castelli, Corinna Ogonowski, Timo Jakobi, Martin Stein, Gunnar Stevens, and Volker Wulf. 2017. What Happened in my home? An end-user development approach for smart home data visualization. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 853–866.

[56] Nico Castelli, Gunnar Stevens, and Timo Jakobi. 2019. Information visualization at home: A literatur e survey of consumption feedback design. (2019).

[57] Spiros Chadoulos, Iordanis Koutsopoulos, and George C Polyzos. 2021. One model fits all: Individualized household energy demand forecasting with a single deep learning model. In *Proceedings of the Twelfth ACM International Conference on Future Energy Systems*. 466–474.

[58] Debaditya Chakraborty, Arafat Alam, Saptarshi Chaudhuri, Hakan Başağaoğlu, Tulio Sulbaran, and Sandeep Langar. 2021. Scenario-based prediction of climate change impacts on building cooling energy consumption with explainable artificial intelligence. *Applied energy* 291 (2021), 116807.

[59] Tanaya Chaudhuri, Yeng Chai Soh, Hua Li, and Lihua Xie. 2019. A feedforward neural network based indoor-climate control framework for thermal comfort and energy saving in buildings. *Applied energy* 248 (2019), 44–53.

[60] Tanaya Chaudhuri, Deqing Zhai, Yeng Chai Soh, Hua Li, and Lihua Xie. 2018. Random forest based thermal comfort prediction from gender-specific physiological parameters using wearable sensing technology. *Energy and Buildings* 166 (2018), 391–406.

[61] Tanaya Chaudhuri, Deqing Zhai, Yeng Chai Soh, Hua Li, and Lihua Xie. 2018. Thermal comfort prediction using normalized skin temperature in a uniform built environment. *Energy and Buildings* 159 (2018), 426–440.

[62] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357. https://doi.org/10.5555/1622407.1622416

[63] Liang Chen, Shuo Xu, Lijun Zhu, Jing Zhang, Xiaoping Lei, and Guancan Yang. 2020. A deep learning based method for extracting semantic information from patent documents. *Scientometrics* 125 (2020), 289–312.

[64] Aniruddh Chennapragada, Divya Periyakoil, Hari Prasanna Das, and Costas J Spanos. 2022. Time series-based deep learning model for personal thermal comfort prediction. In *Proceedings of the Thirteenth ACM International Conference on Future Energy Systems*. 552–555.

[65] Toby C.T. Cheung, Stefano Schiavon, Elliott T. Gall, Ming Jin, and William W. Nazaroff. 2017. Longitudinal assessment of thermal and perceived air quality acceptability in relation to temperature, humidity, and CO2 exposure in Singapore. *Building and Environment* 115 (2017). https://doi.org/10.1016/j.buildenv.2017.01.014

[66] Wonil Choi, Kyungmin Nam, Minwoo Park, Seoyi Yang, Sangyoon Hwang, and Hayoung Oh. 2023. Fake review identification and utility evaluation model using machine learning. *Frontiers in artificial intelligence* 5 (2023), 1064371.

[67] Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz. 2021. I think i get your point, AI! the il-

lusion of explanatory depth in explainable AI. In *26th International Conference on Intelligent User Interfaces*. 307–317.

[68] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555* (2020).

[69] Dave Cliff, Dan Brown, and Philip Treleaven. 2011. Technology trends in the financial markets: A 2020 vision. (2011).

[70] Mark Colley, Benjamin Eder, Jan Ole Rixen, and Enrico Rukzio. 2021. Effects of semantic segmentation visualization on trust, situation awareness, and cognitive load in highly automated vehicles. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–11.

[71] Jonathan Crabbé and Mihaela van der Schaar. [n. d.]. Supplementary Materials For Explaining Time Series Predictions with Dynamic Masks. ([n. d.]).

[72] Jonathan Crabbé and Mihaela Van Der Schaar. 2021. Explaining Time Series Predictions with Dynamic Masks. In *International Conference on Machine Learning*. PMLR, 2166–2177.

[73] Xiaocong Cui, Jung Min Lee, and J Hsieh. 2019. An integrative 3C evaluation framework for explainable artificial intelligence. (2019).

[74] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860* (2019).

[75] Mohammad Dalvi-Esfahani, Mehdi Mosharaf-Dehkordi, Lam Wai Leong, T Ramayah, and Abdulkarim M Jamal Kanaan-Jebna.

2023. Exploring the drivers of XAI-enhanced clinical decision support systems adoption: Insights from a stimulus-organism-response perspective. *Technological Forecasting and Social Change* 195 (2023), 122768.

[76] Devleena Das, Yasutaka Nishimura, Rajan P Vivek, Naoto Takeda, Sean T Fish, Thomas Ploetz, and Sonia Chernova. 2021. Explainable Activity Recognition for Smart Home Systems. *arXiv preprint arXiv:2105.09787* (2021).

[77] Hari Prasanna Das, Stefano Schiavon, and Costas J Spanos. 2021. Unsupervised personal thermal comfort prediction via adversarial domain adaptation. In *Proceedings of the 8th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*. 230–231.

[78] Hari Prasanna Das and Costas J Spanos. 2022. Synthetic personal thermal comfort data generation. In *Proceedings of the 9th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*. 280–281.

[79] Rodrigo Barbosa de Santis, Eduardo Pestana de Aguiar, and Leonardo Goliatt. 2017. Predicting material backorders in inventory management using machine learning. In *2017 IEEE Latin American Conference on Computational Intelligence (LA-CCI)*. IEEE, 1–6. https://doi.org/10.1109/LA-CCI.2017.8285684

[80] N Deshai and B Bhaskara Rao. 2023. Unmasking deception: a CNN and adaptive PSO approach to detecting fake online reviews. *Soft Computing* (2023), 1–22.

[81] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[82] Eva D'hondt, Suzan Verberne, Cornelis Koster, and Lou Boves. 2013. Text representations for patent classification. *Computational Linguistics* 39, 3 (2013), 755–775.

[83] Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. 2018. Explainable artificial intelligence: A survey. In *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*. IEEE, 0210–0215.

[84] Mauro Dragoni and Ivan Donadello. 2022. A knowledge-based strategy for XAI: The explanation graph. *Semantic Web Journal* (2022).

[85] Ramadhani Ally Duma, Zhendong Niu, Ally S Nyamawe, Jude Tchaye-Kondi, and Abdulganiyu Abdu Yusuf. 2023. A Deep Hybrid Model for fake review detection by jointly leveraging review text, overall ratings, and aspect ratings. *Soft Computing* 27, 10 (2023), 6281–6296.

[86] Md Iftekharul Alam Efat, Petr Hajek, Mohammad Zoynul Abedin, Rahat Uddin Azad, Md Al Jaber, Shuvra Aditya, and Mohammad Kabir Hassan. 2022. Deep-learning model using hybrid adaptive trend estimated series for modelling and forecasting sales. *Annals of Operations Research* (2022), 1–32.

[87] Upol Ehsan, Philipp Wintersberger, Q Vera Liao, Martina Mara, Marc Streit, Sandra Wachter, Andreas Riener, and Mark O Riedl. 2021. Operationalizing human-centered perspectives in explainable AI. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, 1–6.

[88] Upol Ehsan, Philipp Wintersberger, Q Vera Liao, Elizabeth Anne Watkins, Carina Manger, Hal Daumé III, Andreas Riener, and Mark O Riedl. 2022. Human-Centered Explainable AI (HCXAI):

beyond opening the black-box of AI. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. ACM, 1–7.

[89] Vijay Ekambaram, Kushagra Manglik, Sumanta Mukherjee, Surya Shravan Kumar Sajja, Satyam Dwivedi, and Vikas Raykar. 2020. Attention based multi-modal new product sales time-series forecasting. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 3110–3118.

[90] Ahmed M Elmogy, Usman Tariq, Mohammed Ammar, and Atef Ibrahim. 2021. Fake reviews detection using supervised machine learning. *International Journal of Advanced Computer Science and Applications* 12, 1 (2021).

[91] Ida Merete Enholm, Emmanouil Papagiannidis, Patrick Mikalef, and John Krogstie. 2022. Artificial intelligence and business value: A literature review. *Information Systems Frontiers* 24, 5 (2022), 1709–1734.

[92] Rocío Escandón, Fabrizio Ascione, Nicola Bianco, Gerardo Maria Mauro, Rafael Suárez, and Juan José Sendra. 2019. Thermal comfort prediction in a building category: Artificial neural network generation from calibrated models for a social housing stock in southern Europe. *Applied Thermal Engineering* 150 (2019), 492–505.

[93] Abinet Tesfaye Eseye and Matti Lehtonen. 2020. Short-term forecasting of heat demand of buildings for efficient and optimal energy management based on integrated machine learning models. *IEEE Transactions on Industrial Informatics* 16, 12 (2020), 7743–7755.

[94] Nasim Eslamirad, Soheil Malekpour Kolbadinejad, Mohammadjavad Mahdavinejad, and Mohammad Mehranrad. 2020. Thermal comfort prediction by applying supervised machine learning

in green sidewalks of Tehran. *Smart and Sustainable Built Environment* 9, 4 (2020), 361–374.

[95] Mohammad Esrafilian-Najafabadi and Fariborz Haghighat. 2022. Impact of predictor variables on the performance of future occupancy prediction: Feature selection using genetic algorithms and machine learning. *Building and Environment* 219 (2022), 109152. https://doi.org/10.1016/j.buildenv.2022.109152

[96] Md Hasib Fakir and Jung Kyung Kim. 2022. Prediction of individual thermal sensation from exhaled breath temperature using a smart face mask. *Building and Environment* 207 (2022), 108507.

[97] Lintao Fang, Le Zhang, Han Wu, Tong Xu, Ding Zhou, and Enhong Chen. 2021. Patent2Vec: Multi-view representation learning on patent-graphs for patent classification. *World Wide Web* 24, 5 (2021), 1791–1812.

[98] Youli Fang, Hong Wang, Lili Zhao, Fengping Yu, and Caiyu Wang. 2020. Dynamic knowledge graph based fake-review detection. *Applied Intelligence* 50 (2020), 4281–4295.

[99] Zahra Qavidel Fard, Zahra Sadat Zomorodian, and Sepideh Sadat Korsavi. 2022. Application of machine learning in thermal comfort studies: A review of methods, performance and challenges. *Energy and Buildings* 256 (2022), 111771.

[100] Yanxiao Feng, Shichao Liu, Julian Wang, Jing Yang, Ying-Ling Jao, and Nan Wang. 2022. Data-driven personal thermal comfort prediction: A literature review. *Renewable and Sustainable Energy Reviews* 161 (2022), 112357.

[101] Luciano Floridi and Massimo Chiriatti. 2020. GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines* 30 (2020), 681–694.

[102] Julien Fontanarava, Gabriella Pasi, and Marco Viviani. 2017. Feature analysis for fake review detection through supervised classification. In *2017 IEEE international conference on data science and advanced Analytics (DSAA)*. IEEE, 658–666.

[103] Maximilian Förster, Mathias Klier, Kilian Kluge, and Irina Sigler. 2020. Fostering human agency: a process for the design of user-centric XAI systems. *International Journal of Intelligent Systems* (2020). https://aisel.aisnet.org/icis2020/hci_artintel/hci_artintel/12

[104] Felix Friedrich, David Steinmann, and Kristian Kersting. 2023. One Explanation Does Not Fit XIL. *arXiv preprint arXiv:2304.07136* (2023).

[105] Haoyun Fu, Styliani Kampezidou, WoongJe Sung, Scott Duncan, and Dimitri N Mavris. 2018. A Data-driven Situational Awareness Approach to Monitoring Campus-wide Power Consumption. In *2018 International Energy Conversion Engineering Conference*. 4414.

[106] Nan Gao, Wei Shao, Mohammad Saiedur Rahaman, Jun Zhai, Klaus David, and Flora D Salim. 2021. Transfer learning for thermal comfort prediction in multiple cities. *Building and Environment* 195 (2021), 107725.

[107] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. 2019. Towards automatic concept-based explanations. *Advances in neural information processing systems* 32 (2019).

[108] Indranil Ghosh, Rabin K Jana, and Mohammad Zoynul Abedin. 2023. An ensemble machine learning framework for Airbnb rental price modeling without using amenity-driven features. *International Journal of Contemporary Hospitality Management* (2023).

[109] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, 80–89.

[110] Brandon M Greenwell, Bradley C Boehmke, and Andrew J McCarthy. 2018. A simple and effective model-based variable importance measure. *arXiv preprint arXiv:1805.04755* (2018).

[111] Ana I Grimaldo and Jasminko Novak. 2019. User-centered visual analytics approach for interactive and explainable energy demand analysis in prosumer scenarios. In *Computer Vision Systems: 12th International Conference, ICVS 2019, Thessaloniki, Greece, September 23–25, 2019, Proceedings 12*. Springer, 700–710.

[112] Ana I Grimaldo and Jasminko Novak. 2020. Combining machine learning with visual analytics for explainable forecasting of energy demand in prosumer scenarios. *Procedia Computer Science* 175 (2020), 525–532.

[113] Phil Grunewald and Marina Diakonova. 2018. The electricity footprint of household activities-implications for demand models. *Energy and Buildings* 174 (2018), 635–641.

[114] Janine Guenther and Oliver Sawodny. 2019. Feature selection and Gaussian Process regression for personalized thermal comfort prediction. *Building and Environment* 148 (2019), 448–458.

[115] Janine Guenther and Oliver Sawodny. 2019. Feature selection for thermal comfort modeling based on constrained LASSO regression. *IFAC-PapersOnLine* 52. Issue 15. `https://doi.org/10.1016/j.ifacol.2019.11.708`

[116] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. 2017. Improved training of wasserstein gans. *Advances in neural information processing systems* 30 (2017).

[117] Karthik S Gurumoorthy, Amit Dhurandhar, Guillermo Cecchi, and Charu Aggarwal. 2019. Efficient data representation by selecting prototypes with importance weights. In *2019 IEEE International Conference on Data Mining (ICDM)*. IEEE, 260–269.

[118] Arousha Haghighian Roudsari, Jafar Afshar, Wookey Lee, and Suan Lee. 2022. PatentNet: multi-label classification of patent documents using deep learning based language understanding. *Scientometrics* (2022), 1–25.

[119] Petr Hajek and Mohammad Zoynul Abedin. 2020. A Profit Function-Maximizing Inventory Backorder Prediction System Using Big Data Analytics. *IEEE Access* 8 (2020), 58982–58994. https://doi.org/10.1109/ACCESS.2020.2983118

[120] Ejaz Ul Haq, Xue Lyu, Youwei Jia, Mengyuan Hua, and Fiaz Ahmad. 2020. Forecasting household electric appliances consumption and peak demand based on hybrid machine learning approach. *Energy Reports* 6 (2020), 1099–1105.

[121] AKM Bahalul Haque, AKM Najmul Islam, and Patrick Mikalef. 2023. Explainable Artificial Intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research. *Technological Forecasting and Social Change* 186 (2023), 122120.

[122] Peter Hase and Mohit Bansal. 2020. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? *arXiv preprint arXiv:2005.01831* (2020).

[123] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*. IEEE, 1322–1328. https://doi.org/10.1109/IJCNN.2008.4633969

[124] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015).

[125] Robert R Hoffman, Shane T Mueller, and Gary Klein. 2017. Explaining explanation, part 2: Empirical foundations. *IEEE Intelligent Systems* 32, 4 (2017), 78–86.

[126] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608* (2018).

[127] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. 2019. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9, 4 (2019), e1312.

[128] Bingquan Huang, Mohand Tahar Kechadi, and Brian Buckley. 2012. Customer churn prediction in telecommunications. *Expert Systems with Applications* 39, 1 (2012), 1414–1425.

[129] Walayat Hussain, Honghao Gao, Muhammad Raheel Raza, Fethi A Rabhi, and Jose M Merigo. 2022. Assessing cloud QoS predictions using OWA in neural network methods. *Neural Computing and Applications* (2022), 1–18. https://doi.org/10.1007/s00521-022-07297-z

[130] Walayat Hussain, José M Merigó, and Muhammad Raheel Raza. 2021. Predictive intelligence using ANFIS-induced OWAWA for

complex stock market prediction. *International Journal of Intelligent Systems* (2021). https://doi.org/10.1002/int.22732

[131] Walayat Hussain, José M Merigó, Muhammad Raheel Raza, and Honghao Gao. 2022. A new QoS prediction model using hybrid IOWA-ANFIS with fuzzy C-means, subtractive clustering and grid partitioning. *Information Sciences* 584 (2022), 280–300. https://doi.org/10.1016/j.ins.2021.10.054

[132] Hilary Hutchinson, Wendy Mackay, Bo Westerlund, Benjamin B Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane Conversy, Helen Evans, Heiko Hansen, et al. 2003. Technology probes: inspiring design for and with families. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 17–24.

[133] Igor Ilic, Berk Görgülü, Mucahit Cevik, and Mustafa Gökçe Baydoğan. 2021. Explainable boosted linear regression for time series forecasting. *Pattern Recognition* 120 (2021), 108144.

[134] Samiul Islam and Saman Hassanzadeh Amin. 2020. Prediction of probable backorder scenarios in the supply chain using Distributed Random Forest and Gradient Boosting Machine learning techniques. *Journal of Big Data* 7, 1 (2020), 1–22. https://doi.org/10.1186/s40537-020-00345-2

[135] Sukirty Jain, Sanyam Shukla, and Rajesh Wadhvani. 2018. Dynamic selection of normalization techniques using data complexity measures. *Expert Systems with Applications* 106 (2018), 252–262.

[136] Timo Jakobi, Gunnar Stevens, Nico Castelli, Corinna Ogonowski, Florian Schaub, Nils Vindice, Dave Randall, Peter Tolmie, and Volker Wulf. 2018. Evolving needs in IoT control and accountability: A longitudinal study on smart home intelligibility. *Proceedings*

of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 2, 4 (2018), 1–28.

[137] Christian Janiesch, Patrick Zschech, and Kai Heinrich. 2021. Machine learning and deep learning. *Electronic Markets* 31, 3 (2021), 685–695. https://doi.org/10.1007/s12525-021-00475-2

[138] Nathalie Japkowicz and Mohak Shah. 2011. *Evaluating learning algorithms: a classification perspective*. Cambridge University Press.

[139] Prageeth Jayathissa, Matias Quintana, Mahmoud Abdelrahman, and Clayton Miller. 2020. Humans-as-a-sensor for buildings—intensive longitudinal indoor comfort models. *Buildings* 10, 10 (2020), 174.

[140] Rikke Hagensby Jensen, Yolande Strengers, Jesper Kjeldskov, Larissa Nicholls, and Mikael B Skov. 2018. Designing the desirable smart home: A study of household experiences and energy consumption impacts. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–14.

[141] Shuo Jiang, Jie Hu, Christopher L Magee, and Jianxi Luo. 2022. Deep learning for technical document classification. *IEEE Transactions on Engineering Management* (2022).

[142] Jiao Jiao, Heike Brugger, Michael Behrisch, and Wolfgang Eichhammer. 2022. Identifying drivers of residential energy consumption by explainable energy demand forecasting. (2022).

[143] Licheng Jiao and Jin Zhao. 2019. A survey on the new generation of deep learning in image processing. *IEEE Access* 7 (2019), 172231–172263. https://doi.org/10.1109/ACCESS.2019.2956508

[144] Chen Jing-Yu and Wang Ya-Jun. 2022. Semi-supervised fake re-

views detection based on aspamgan. *Journal of Artificial Intelligence* 4, 1 (2022), 17–36.

[145] M Humayn Kabir, Khondokar Fida Hasan, Mohammad Kamrul Hasan, and Keyvan Ansari. 2021. Explainable Artificial Intelligence for Smart City Application: A Secure and Trusted Platform. *arXiv preprint arXiv:2111.00601* (2021).

[146] Eleni Kamateri, Vasileios Stamatis, Konstantinos Diamantaras, and Michail Salampasis. 2022. Automated Single-Label Patent Classification using Ensemble Classifiers. In *2022 14th International Conference on Machine Learning and Computing (ICMLC)*. 324–330.

[147] Uday Kamath and John Liu. 2021. *Explainable artificial intelligence: an introduction to interpretable machine learning*. Vol. 2. Springer.

[148] Myungchul Kang, Suan Lee, and Wookey Lee. 2020. Prior art search using multi-modal embedding of patent documents. In *2020 IEEE International Conference on Big Data and Smart Computing (BigComp)*. IEEE, 548–550.

[149] Md Rezaul Karim, Sumon Kanti Dey, Tanhim Islam, Sagor Sarker, Mehadi Hasan Menon, Kabir Hossain, Md Azam Hossain, and Stefan Decker. 2021. Deephateexplainer: Explainable hate speech detection in under-resourced bengali language. In *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 1–10.

[150] Md Rezaul Karim, Tanhim Islam, Md Shajalal, Oya Beyan, Christoph Lange, Michael Cochez, Dietrich Rebholz-Schuhmann, and Stefan Decker. 2023. Explainable AI for Bioinformatics: Methods, Tools and Applications. *Briefings in bioinformatics* 24, 5 (2023), bbad236.

[151] Md Rezaul Karim, Md Shajalal, Alexander Graß, Till Döhmen, Sisay Adugna Chala, Alexander Boden, Christian Beecks, and Stefan Decker. 2023. Interpreting black-box machine learning models for high dimensional datasets. In *2023 IEEE 10th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 1–10.

[152] Katarina Katić, Rongling Li, and Wim Zeiler. 2020. Machine learning algorithms applied to a prediction of personal overall thermal comfort using skin temperatures and occupants' heating behavior. *Applied Ergonomics* 85 (2020), 103078. https://doi.org/10.1016/j.apergo.2020.103078

[153] Mohammad-Rasool Kazemzadeh, Ali Amjadian, and Turaj Amraee. 2020. A hybrid data mining driven algorithm for long term electric peak load and energy demand forecasting. *Energy* 204 (2020), 117948.

[154] Willett Kempton. 1986. Two theories of home heat control. *Cognitive science* 10, 1 (1986), 75–90.

[155] Elham Khodabandehloo, Daniele Riboni, and Abbas Alimohammadi. 2021. HealthXAI: Collaborative and explainable AI for supporting early diagnosis of cognitive decline. *Future Generation Computer Systems* 116 (2021), 168–189.

[156] Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. 2016. Examples are not enough, learn to criticize! criticism for interpretability. *Advances in neural information processing systems* 29 (2016).

[157] Buomsoo Raymond Kim, Karthik Srinivasan, Sung Hye Kong, Jung Hee Kim, Chan Soo Shin, and Sudha Ram. 2023. ROLEX: A NOVEL METHOD FOR INTERPRETABLE MACHINE LEARNING USING ROBUST LOCAL EXPLANATIONS. *MIS Quarterly* 47, 3 (2023).

[158] Doha Kim, Yeosol Song, Songyie Kim, Sewang Lee, Yanqin Wu, Jungwoo Shin, and Daeho Lee. 2023. How should the results of artificial intelligence be explained to users?-Research on consumer preferences in user-centered explainable artificial intelligence. *Technological Forecasting and Social Change* 188 (2023), 122343.

[159] Dae Hyun Kim, Enamul Hoque, and Maneesh Agrawala. 2020. Answering questions about charts and generating visual explanations. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13.

[160] Juram Kim, Gyumin Lee, Seungbin Lee, and Changyong Lee. 2022. Towards expert–machine collaborations for technology valuation: An interpretable machine learning approach. *Technological Forecasting and Social Change* 183 (2022), 121940.

[161] Jin-Young Kim and Sung-Bae Cho. 2019. Electric energy consumption prediction by deep learning with state explainable autoencoder. *Energies* 12, 4 (2019), 739.

[162] Jin-Young Kim and Sung-Bae Cho. 2020. Electric Energy Demand Forecasting with Explainable Time-series Modeling. In *2020 International Conference on Data Mining Workshops (ICDMW)*. IEEE, 711–716.

[163] Jin-Young Kim and Sung-Bae Cho. 2021. Explainable prediction of electric energy demand using a deep autoencoder with interpretable latent space. *Expert Systems with Applications* 186 (2021), 115842.

[164] Tae-Young Kim and Sung-Bae Cho. 2019. Predicting residential energy consumption using CNN-LSTM neural networks. *Energy* 182 (2019), 72–81.

[165] Durk P Kingma, Tim Salimans, and Max Welling. 2015. Variational dropout and the local reparameterization trick. *Advances in neural information processing systems* 28 (2015). https://doi.org/10.48550/arXiv.1506.02557

[166] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, et al. 2020. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896* (2020).

[167] Michal Kucer, Diane Oyen, Juan Castorena, and Jian Wu. 2022. DeepPatent: Large scale patent drawing recognition and retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2309–2318.

[168] Lenneke Kuijer and Elisa Giaccardi. 2018. Co-performance: Conceptualizing the role of artificial agency in the design of everyday life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.

[169] Rahul Kumar, Shubhadeep Mukherjee, and Nripendra P Rana. 2023. Exploring Latent Characteristics of Fake Reviews and Their Intermediary Role in Persuading Buying Decisions. *Information Systems Frontiers* (2023), 1–18.

[170] Dattatray Vishnu Kute, Biswajeet Pradhan, Nagesh Shukla, and Abdullah Alamri. 2021. Deep learning and explainable artificial intelligence techniques applied for detecting money laundering–a critical review. *IEEE Access* 9 (2021), 82300–82317.

[171] Piotr Ładyżyński, Kamil Żbikowski, and Piotr Gawrysiak. 2019. Direct marketing campaigns in retail banking with the use of deep learning and random forests. *Expert Systems with Applications* 134 (2019), 28–35. https://doi.org/10.1016/j.eswa.2019.05.020

[172] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. 2019. Faithful and customizable explanations of black box models. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 131–138.

[173] Himabindu Lakkaraju, Dylan Slack, Yuxin Chen, Chenhao Tan, and Sameer Singh. 2022. Rethinking explainability as a dialogue: A practitioner's perspective. *arXiv preprint arXiv:2202.01875* (2022).

[174] SO LAWAL and KG AKINTOLA. 2021. A Product Backorder Predictive Model Using Recurrent Neural Network. *IRE Journals* 4, 8 (2021).

[175] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*. PMLR, 1188–1196.

[176] Freddy Lecue. 2020. On the role of knowledge graphs in explainable AI. *Semantic Web* 11, 1 (2020), 41–51.

[177] Jeehee Lee and Youngjib Ham. 2021. Physiological sensing-driven personal thermal comfort modelling in consideration of human activity variations. *Building Research and Information* 49 (2021), 512–524. Issue 5. https://doi.org/10.1080/09613218.2020.1840328

[178] Jieh-Sheng Lee and Jieh Hsiang. 2020. Patent classification by fine-tuning BERT language model. *World Patent Information* 61 (2020), 101965.

[179] Seungjae Lee, Panagiota Karava, Athanasios Tzempelikos, and Ilias Bilionis. 2020. A smart and less intrusive feedback request algorithm towards human-centered HVAC operation. *Building and*

Environment 184 (10 2020), 107190. https://doi.org/10.1016/j.buildenv.2020.107190

[180] Da Li, Carol C. Menassa, Vineet R. Kamat, and Eunshin Byon. 2020. HEAT - Human Embodied Autonomous Thermostat. *Building and Environment* 178 (2020). https://doi.org/10.1016/j.buildenv.2020.106879

[181] Guannan Li, Yubei Wu, Jiangyan Liu, Xi Fang, and Zixi Wang. 2022. Performance evaluation of short-term cross-building energy predictions using deep transfer learning strategies. *Energy and Buildings* 275 (2022), 112461.

[182] Huahang Li, Shuangyin Li, Yuncheng Jiang, and Gansen Zhao. 2022. CoPatE: A Novel Contrastive Learning Framework for Patent Embeddings. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 1104–1113.

[183] Shaobo Li, Jie Hu, Yuxin Cui, and Jianjun Hu. 2018. DeepPatent: patent classification with convolutional neural networks and word embedding. *Scientometrics* 117 (2018), 721–744.

[184] Tong Li, Zhaohua Wang, and Wenhui Zhao. 2022. Comparison and application potential analysis of autoencoder-based electricity pattern mining algorithms for large-scale demand response. *Technological Forecasting and Social Change* 177 (2022), 121523.

[185] Yuqi Li. 2017. Backorder prediction using machine learning for Danish craft beer breweries. *PhD diss., Aalborg University* (2017).

[186] Yu Li, Chao Huang, Lizhong Ding, Zhongxiao Li, Yijie Pan, and Xin Gao. 2019. Deep learning in bioinformatics: Introduction, application, and perspective in the big data era. *Methods* 166 (2019), 4–21. https://doi.org/10.1016/j.ymeth.2019.04.008

[187] Meng Qi Liao. 2020. How should AI talk to users when collecting their personal information? Effects of role framing and self-referencing on Human-AI interaction. (2020).

[188] Q Vera Liao and Kush R Varshney. 2021. Human-centered explainable ai (xai): From algorithms to user experiences. *arXiv preprint arXiv:2110.10790* (2021).

[189] Zachary C Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57.

[190] Kuixing Liu, Ting Nie, Wei Liu, Yiqing Liu, and Dayi Lai. 2020. A machine learning approach to predict outdoor thermal comfort using local skin temperatures. *Sustainable Cities and Society* 59 (2020), 102216.

[191] Shichao Liu, Stefano Schiavon, Hari Prasanna Das, Ming Jin, and Costas J Spanos. 2019. Personal thermal comfort models with wearable sensors. *Building and Environment* 162 (2019), 106281.

[192] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[193] Sandra Maria Correia Loureiro, João Guerreiro, and Iis Tussyadiah. 2021. Artificial intelligence in business: State of the art and future research agenda. *Journal of business research* 129 (2021), 911–926.

[194] Siliang Lu, Weilong Wang, Shihan Wang, and Erica Cochran Hameen. 2019. Thermal comfort-based personalized models with non-intrusive sensing technique in office buildings. *Applied Sci-*

ences (Switzerland) 9 (2019). Issue 9. https://doi.org/10.3390/app9091768

[195] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).

[196] Mengzhen Luo, Xiaoyu Shi, Qianqian Ji, Mingsheng Shang, Xianbo He, and Weiguo Tao. 2020. A Deep Self-learning Classification Framework for Incomplete Medical Patents with Multi-label. In *Advances in Natural Computation, Fuzzy Systems and Knowledge Discovery: Volume 2*. Springer, 566–573.

[197] Yunlong Ma, Xiao Chen, Liming Wang, and Jianlan Yang. 2021. Study on Smart Home Energy Management System Based on Artificial Intelligence. *Journal of Sensors* (2021).

[198] Larissa Arakawa Martins, Veronica Soebarto, and Terence Williamson. 2022. A systematic review of personal thermal comfort models. *Building and Environment* 207 (2022), 108502.

[199] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 14867–14875.

[200] Winston Maxwell and Bruno Dumas. 2023. Meaningful XAI based on user-centric design methodology. *arXiv preprint arXiv:2308.13228* (2023).

[201] Ross May, Xingxing Zhang, Jinshun Wu, and Mengjie Han. 2019. Reinforcement learning control for indoor comfort: a survey. In *IOP Conference Series: Materials Science and Engineering*, Vol. 609. IOP Publishing, 062011.

[202] Christian Meske, Enrico Bunde, Johannes Schneider, and Martin Gersch. 2022. Explainable artificial intelligence: objectives, stakeholders, and future research opportunities. *Information Systems Management* 39, 1 (2022), 53–63. https://doi.org/10.1080/10580530.2020.1849465

[203] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems* 26 (2013).

[204] Hokey Min. 2010. Artificial intelligence in supply chain management: theory and applications. *International Journal of Logistics: Research and Applications* 13, 1 (2010), 13–39.

[205] Abrar Qadir Mir, Furqan Yaqub Khan, and Mohammad Ahsan Chishti. 2023. Online Fake Review Detection Using Supervised Machine Learning And Bert Model. *arXiv preprint arXiv:2301.03225* (2023).

[206] Rami Mohawesh, Son Tran, Robert Ollington, and Shuxiang Xu. 2021. Analysis of concept drift in fake reviews detection. *Expert Systems with Applications* 169 (2021), 114318.

[207] Rami Mohawesh, Shuxiang Xu, Matthew Springer, Yaser Jararweh, Muna Al-Hawawreh, and Sumbal Maqsood. 2023. An Explainable Ensemble of Multi-View Deep Learning Model for Fake Review Detection. *Journal of King Saud University-Computer and Information Sciences* (2023), 101644.

[208] Rami Mohawesh, Shuxiang Xu, Son N Tran, Robert Ollington, Matthew Springer, Yaser Jararweh, and Sumbal Maqsood. 2021. Fake reviews detection: A survey. *IEEE Access* 9 (2021), 65771–65802.

[209] Christoph Molnar. 2020. *Interpretable machine learning*. Lulu. com.

[210] Grégoire Montavon. 2019. Gradient-based vs. propagation-based explanations: An axiomatic comparison. *Explainable ai: Interpreting, explaining and visualizing deep learning* (2019), 253–265.

[211] Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. 2019. Layer-wise relevance propagation: an overview. *Explainable AI: interpreting, explaining and visualizing deep learning* (2019), 193–209.

[212] Vincenzo Moscato, Antonio Picariello, and Giancarlo Sperlí. 2021. A benchmark of machine learning approaches for credit score prediction. *Expert Systems with Applications* 165 (2021), 113986. https://doi.org/10.1016/j.eswa.2020.113986

[213] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 607–617. https://doi.org/10.1145/3351095.3372850

[214] Haider Mshali, Tayeb Lemlouma, Maria Moloney, and Damien Magoni. 2018. A survey on health monitoring systems for health smart homes. *International Journal of Industrial Ergonomics* 66 (2018), 26–56.

[215] Henrik Mucha, Sebastian Robert, Rüdiger Breitschwerdt, and Michael Fellmann. 2020. Towards participatory design spaces for explainable ai interfaces in expert domains. In *43rd German Conference on Artificial Intelligence, Bamberg, Germany*.

[216] Henrik Mucha, Sebastian Robert, Ruediger Breitschwerdt, and Michael Fellmann. 2021. Interfaces for Explanations in Human-

AI Interaction: Proposing a Design Evaluation Approach. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–6.

[217] David Murray, Lina Stankovic, and Vladimir Stankovic. 2017. An electrical load measurements dataset of United Kingdom households from a two-year longitudinal study. *Scientific data* 4, 1 (2017), 1–12.

[218] Alexandre Moreira Nascimento, Lucio Flavio Vismari, Caroline Bianca Santos Tancredi Molina, Paulo Sergio Cugnasca, Joao Batista Camargo, Jorge Rady de Almeida, Rafia Inam, Elena Fersman, Maria Valeria Marquezini, and Alberto Yukinobu Hata. 2019. A systematic literature review about the impact of artificial intelligence on autonomous vehicle safety. *IEEE Transactions on Intelligent Transportation Systems* 21, 12 (2019), 4928–4946.

[219] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice Van Keulen, and Christin Seifert. 2023. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *Comput. Surveys* 55, 13s (2023), 1–42.

[220] Jack Ngarambe, Geun Young Yun, and Mat Santamouris. 2020. The use of artificial intelligence (AI) methods in the prediction of thermal comfort in buildings: Energy implications of AI-based thermal comfort controls. *Energy and Buildings* 211 (2020), 109807.

[221] Jakob Nielsen and Rolf Molich. 1990. Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 249–256.

[222] Tuomas Nissinen. 2015. User experience prototyping: a literature review. (2015).

[223] Mahsan Nourani, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrima Rahman, Eric Ragan, and Vibhav Gogate. 2021. Anchoring bias affects mental model formation and user reliance in explainable ai systems. In *26th International Conference on Intelligent User Interfaces*. 340–350.

[224] Andrzej S Nowak and Tadeusz Radzik. 1994. The Shapley value for n-person games in generalized characteristic function form. *Games and Economic Behavior* 6, 1 (1994), 150–161. https://doi.org/10.1006/game.1994.1008

[225] Charis Ntakolia, Christos Kokkotiis, Serafeim Moustakidis, and Elpiniki Papageorgiou. 2021. An explainable machine learning pipeline for backorder prediction in inventory management systems. In *25th Pan-Hellenic Conference on Informatics*. 229–234. https://doi.org/10.1145/3503823.3503866

[226] Charis Ntakolia, Christos Kokkotis, Patrik Karlsson, and Serafeim Moustakidis. 2021. An Explainable Machine Learning Model for Material Backorder Prediction in Inventory Management. *Sensors* 21, 23 (2021), 7926. https://doi.org/10.3390/s21237926

[227] Jeroen Ooge, Shotallo Kato, and Katrien Verbert. 2022. Explaining recommendations in e-learning: Effects on adolescents' trust. In *27th International Conference on Intelligent User Interfaces*. 93–105.

[228] Arjun Panesar. 2019. *Machine learning and AI for healthcare*. Springer. 1–407 pages. https://doi.org/10.1007/978-1-4842-3799-1

[229] Cecilia Panigutti, Andrea Beretta, Fosca Giannotti, and Dino Pedreschi. 2022. Understanding the impact of explanations on advice-taking: a user study for AI-based clinical Decision Sup-

port Systems. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–9.

[230] Nidhi A Patel and Rakesh Patel. 2018. A survey on fake review detection using machine learning techniques. In *2018 4th international Conference on computing Communication and automation (ICCCA)*. IEEE, 1–6.

[231] Himangshu Paul and Alexander Nikolaev. 2021. Fake review detection on online E-commerce platforms: a systematic literature review. *Data Mining and Knowledge Discovery* 35, 5 (2021), 1830–1881.

[232] Shiliang Peng, Lin Fan, Li Zhang, Huai Su, Yuxuan He, Qian He, Xiao Wang, Dejun Yu, and Jinjun Zhang. 2024. Spatio-temporal prediction of total energy consumption in multiple regions using explainable deep neural network. *Energy* (2024), 131526.

[233] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[234] Maria Petrescu, Haya Ajjan, and Dana L Harrison. 2023. Man vs machine–Detecting deception in online reviews. *Journal of Business Research* 154 (2023), 113346.

[235] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–52.

[236] Subhash Chandra Pujari, Annemarie Friedrich, and Jannik Strötgen. 2021. A multi-task approach to neural multi-label hierarchi-

cal patent classification using transformers. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part I 43*. Springer, 513–528.

[237] Matias Quintana, Stefano Schiavon, Kwok Wai Tham, and Clayton Miller. 2020. Balancing thermal comfort datasets: We GAN, but should we?. In *Proceedings of the 7th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*. 120–129.

[238] Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations as mechanisms for supporting algorithmic transparency. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–13.

[239] Arun Rai. 2020. Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science* 48, 1 (2020), 137–141.

[240] Prajit Ramachandran, Barret Zoph, and Quoc V Le. 2017. Searching for activation functions. *arXiv preprint arXiv:1710.05941* (2017). https://doi.org/10.48550/arXiv.1710.05941

[241] Sahana Ramnath, Preksha Nema, Deep Sahni, and Mitesh M Khapra. 2020. Towards interpreting BERT for reading comprehension based QA. *arXiv preprint arXiv:2010.08983* (2020).

[242] Kuldeep Randhawa, Chu Kiong Loo, Manjeevan Seera, Chee Peng Lim, and Asoke K Nandi. 2018. Credit card fraud detection using AdaBoost and majority voting. *IEEE access* 6 (2018), 14277–14284. https://doi.org/10.1109/ACCESS.2018.2806420

[243] Kaivalya Rawal and Himabindu Lakkaraju. 2020. Beyond individualized recourse: Interpretable and interactive summaries of

actionable recourses. *Advances in Neural Information Processing Systems* 33 (2020), 12187–12198.

[244] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " 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*. 1135–1144.

[245] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386* (2016).

[246] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[247] Daniele Riboni. 2021. Keynote: Explainable AI in Pervasive Healthcare: Open Challenges and Research Directions. In *2021 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*. IEEE, 1–1.

[248] Markus Rohde, Peter Brödner, Gunnar Stevens, Matthias Betz, and Volker Wulf. 2017. Grounded Design–a praxeological IS research perspective. *Journal of Information Technology* 32 (2017), 163–179.

[249] Thomas Rojat, Raphaël Puget, David Filliat, Javier Del Ser, Rodolphe Gelin, and Natalia Díaz-Rodríguez. 2021. Explainable artificial intelligence (xai) on timeseries data: A survey. *arXiv preprint arXiv:2104.00950* (2021).

[250] Yao Rong, Tobias Leemann, Thai-Trang Nguyen, Lisa Fiedler, Peizhu Qian, Vaibhav Unhelkar, Tina Seidel, Gjergji Kasneci, and Enkelejda Kasneci. 2023. Towards human-centered explainable

ai: A survey of user studies for model explanations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).

[251] Yao Rong, Tobias Leemann, Thai-trang Nguyen, Lisa Fiedler, Tina Seidel, Gjergji Kasneci, and Enkelejda Kasneci. 2022. Towards Human-centered Explainable AI: User Studies for Model Explanations. *arXiv preprint arXiv:2210.11584* (2022).

[252] Arousha Haghighian Roudsari, Jafar Afshar, Charles Cheolgi Lee, and Wookey Lee. 2020. Multi-label patent classification using attention-aware deep learning model. In *2020 IEEE International Conference on Big Data and Smart Computing (BigComp)*. IEEE, 558–559.

[253] Arousha Haghighian Roudsari, Jafar Afshar, Suan Lee, and Wookey Lee. 2021. Comparison and analysis of embedding methods for patent documents. In *2021 IEEE International Conference on Big Data and Smart Computing (BigComp)*. IEEE, 152–155.

[254] Jitendra Kumar Rout, Amiya Kumar Dash, and Niranjan Kumar Ray. 2018. A framework for fake review detection: issues and challenges. In *2018 international conference on information technology (ICIT)*. IEEE, 7–10.

[255] Joze M Rozanec. 2021. Explainable demand forecasting: A data mining goldmine. In *Companion Proceedings of the Web Conference 2021*. 723–724.

[256] Amal Saadallah, Matthias Jakobs, and Katharina Morik. 2021. Explainable online deep neural network selection using adaptive saliency maps for time series forecasting. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 404–420.

[257] Amal Saadallah, Matthias Jakobs, and Katharina Morik. 2022.

Explainable online ensemble of deep neural network pruning for time series forecasting. *Machine Learning* (2022), 1–29.

[258] Nikos D Sakkas, Sofia Yfanti, Pooja Shah, Nikitas Sakkas, Christina Chaniotakis, Costas Daskalakis, Eduard Barbu, and Marharyta Domnich. 2023. Explainable Approaches for Forecasting Building Electricity Consumption. (2023).

[259] Joni Salminen, Chandrashekhar Kandpal, Ahmed Mohamed Kamel, Soon-gyo Jung, and Bernard J Jansen. 2022. Creating and detecting fake reviews of online products. *Journal of Retailing and Consumer Services* 64 (2022), 102771.

[260] Rohit Saluja, Avleen Malhi, Samanta Knapič, Kary Främling, and Cicek Cavdar. 2021. Towards a rigorous evaluation of explainability for multivariate time series. *arXiv preprint arXiv:2104.04075* (2021).

[261] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).

[262] Gunjan Saraogi, Deepa Gupta, Lavanya Sharma, and Ajay Rana. 2021. An Un-Supervised Approach for Backorder Prediction Using Deep Autoencoder. *Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science)* 14, 2 (2021), 500–511. http://dx.doi.org/10.2174/2213275912666190819112609

[263] Vitaly Schetinin, Jonathan E Fieldsend, Derek Partridge, Timothy J Coats, Wojtek J Krzanowski, Richard M Everson, Trevor C Bailey, and Adolfo Hernandez. 2007. Confident interpretation of Bayesian decision tree ensembles for clinical applications. *IEEE Transactions on Information Technology in Biomedicine* 11, 3 (2007), 312–319.

[264] Udo Schlegel, Daniela Oelke, Daniel A Keim, and Mennatallah El-Assady. 2020. An empirical study of explainable AI techniques on deep learning models for time series tasks. *arXiv preprint arXiv:2012.04344* (2020).

[265] Jakob Schoeffer and Niklas Kuehl. 2021. Appropriate fairness perceptions? On the effectiveness of explanations in enabling people to assess the fairness of automated decision systems. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*. 153–157.

[266] Jakob Schoeffer, Niklas Kuehl, and Yvette Machowski. 2022. "There is not enough information": On the effects of explanations on perceptions of informational fairness and trustworthiness in automated decision-making. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1616–1628.

[267] Tobias Schwartz, Sebastian Denef, Gunnar Stevens, Leonardo Ramirez, and Volker Wulf. 2013. Cultivating energy literacy: results from a longitudinal living lab study of a home energy management system. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1193–1202.

[268] Tobias Schwartz, Sebastian Denef, Gunnar Stevens, Leonardo Ramirez, and Volker Wulf. 2013. Cultivating energy literacy: results from a longitudinal living lab study of a home energy management system. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1193–1202.

[269] Tobias Schwartz, Gunnar Stevens, Leonardo Ramirez, and Volker Wulf. 2013. Uncovering practices of making energy consumption accountable: A phenomenological inquiry. *ACM Transactions on Computer-Human Interaction (TOCHI)* 20, 2 (2013), 1–30.

[270] Tobias Schwartz, Gunnar Stevens, Leonardo Ramirez, and Volker Wulf. 2013. Uncovering practices of making energy consumption accountable: A phenomenological inquiry. *ACM Transactions on Computer-Human Interaction (TOCHI)* 20, 2 (2013), 1–30.

[271] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.

[272] Neil Selwyn. 2022. The future of AI and education: Some cautionary notes. *European Journal of Education* 57, 4 (2022), 620–631.

[273] Md Shajalal, Mohammad Zoynul Abedin, and Mohammed Mohi Uddin. [n. d.]. Handling class imbalance data in business domain. In *The Essentials of Machine Learning in Finance and Accounting*. Routledge, 199–210.

[274] Md Shajalal and Masaki Aono. 2018. Sentence-level semantic textual similarity using word-level semantics. In *2018 10th International Conference on Electrical and Computer Engineering (ICECE)*. IEEE, 113–116.

[275] Md Shajalal and Masaki Aono. 2019. Semantic textual similarity between sentences using bilingual word semantics. *Progress in Artificial Intelligence* 8 (2019), 263–272.

[276] Md Shajalal and Masaki Aono. 2020. Coverage-based query subtopic diversification leveraging semantic relevance. *Knowledge and Information Systems* 62 (2020), 2873–2891.

[277] Md Shajalal, Alexander Boden, and Gunnar Stevens. 2022. Explainable product backorder prediction exploiting CNN: Introduc-

ing explainable models in businesses. *Electronic Markets* (2022), 1–16.

[278] Md Shajalal, Alexander Boden, and Gunnar Stevens. 2022. Towards user-centered explainable energy demand forecasting systems. In *Proceedings of the Thirteenth ACM International Conference on Future Energy Systems*. 446–447.

[279] Md Shajalal, Alexander Boden, and Gunnar Stevens. 2024. ForecastExplainer: Explainable household energy demand forecasting by approximating shapley values using DeepLIFT. *Technological Forecasting and Social Change* 206 (2024), 123588.

[280] Md Shajalal, Alexander Boden, Gunnar Stevens, Delong Du, and Dean-Robin Kern. 2024. Explaining AI Decisions: Towards Achieving Human-Centered Explainability in Smart Home Environments. *arXiv preprint arXiv:2404.16074* (2024).

[281] Md Shajalal, Milad Bohlouli, Hari Prasanna Das, Alexander Boden, and Gunnar Stevens. 2022. Focus on What Matters: Improved Feature Selection Techniques for Personal Thermal Comfort Modeling. In *Proceedings of the The 9th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (BuildSys 22)*. ACM.

[282] Md Shajalal, Milad Bohlouli, Hari Prasanna Das, Alexander Boden, and Gunnar Stevens. 2022. Focus on what matters: improved feature selection techniques for personal thermal comfort modelling. In *Proceedings of the 9th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*. 496–499.

[283] Md Shajalal, Milad Bohlouli, Hari Prasanna Das, Alexander Boden, and Gunnar Stevens. 2024. Improved Thermal Comfort

Model Leveraging Conditional Tabular GAN Focusing on Feature Selection. *IEEE Access* 12 (2024), 30039–30053.

[284] Md Shajalal, Sebastian Denef, Md. Rezaul Karim, Alexander Boden, and Stevens Gunnar. 2023. Unveiling Black-boxes: Explainable Deep Learning Models for Patent Classification. In *The World Conference on eXplainable Artificial Intelligence 2023*. Springer, 1–18.

[285] Md Shajalal, Sebastian Denef, Md Rezaul Karim, Alexander Boden, and Gunnar Stevens. 2023. Unveiling Black-Boxes: Explainable Deep Learning Models for Patent Classification. In *World Conference on Explainable Artificial Intelligence*. Springer, 457–474.

[286] Md Shajalal, Petr Hajek, and Mohammad Zoynul Abedin. 2021. Product backorder prediction using deep neural network on imbalanced data. *International Journal of Production Research* (2021), 1–18. https://doi.org/10.1080/00207543.2021.1901153

[287] Md Shajalal, Petr Hajek, and Mohammad Zoynul Abedin. 2023. Product backorder prediction using deep neural network on imbalanced data. *International Journal of Production Research* 61, 1 (2023), 302–319.

[288] Marawan Shalaby, Jan Stutzki, Matthias Schubert, and Stephan Günnemann. 2018. An lstm approach to patent classification based on fixed hierarchy vectors. In *Proceedings of the 2018 SIAM International Conference on Data Mining*. SIAM, 495–503.

[289] Chengcheng Shan, Jiawen Hu, Jianhong Wu, Aili Zhang, Guoliang Ding, and Lisa X. Xu. 2020. Towards non-intrusive and high accuracy prediction of personal thermal comfort using a few sensitive physiological parameters. *Energy and Buildings* 207 (2020), 109594. https://doi.org/10.1016/j.enbuild.2019.109594

[290] Guohou Shan, Lina Zhou, and Dongsong Zhang. 2021. From conflicts and confusion to doubts: Examining review inconsistency for fake review detection. *Decision Support Systems* 144 (2021), 113513.

[291] Mike Shann, Alper Alan, Sven Seuken, Enrico Costanza, and Sarvapali D Ramchurn. 2017. Save money or feel cozy?: A field experiment evaluation of a smart thermostat that learns heating preferences. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, Vol. 16. International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS).

[292] Eva Sharma, Chen Li, and Lu Wang. 2019. BIGPATENT: A large-scale dataset for abstractive and coherent summarization. *arXiv preprint arXiv:1906.03741* (2019).

[293] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *International conference on machine learning*. PMLR, 3145–3153.

[294] Zhang Shunxiang, Zhu Aoqiang, Zhu Guangli, Wei Zhongliang, and Li KuanChing. 2023. Building fake review detection model based on sentiment intensity and PU learning. *IEEE Transactions on Neural Networks and Learning Systems* (2023).

[295] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013).

[296] Digvijay Singh, Minakshi Memoria, and Rajiv Kumar. 2023. Deep Learning Based Model for Fake Review Detection. In *2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT)*. IEEE, 92–95.

[297] Rahul Singhal and Rasha Kashef. 2023. A Weighted Stacking Ensemble Model With Sampling for Fake Reviews Detection. *IEEE Transactions on Computational Social Systems* (2023).

[298] Leon Sixt, Martin Schuessler, Oana-Iuliana Popescu, Philipp Weiß, and Tim Landgraf. 2022. Do users benefit from interpretable vision? a user study, baseline, and dataset. *arXiv preprint arXiv:2204.11642* (2022).

[299] Nivethitha Somu, Anirudh Sriram, Anupama Kowli, and Krithi Ramamritham. 2021. A hybrid deep transfer learning strategy for thermal comfort prediction in buildings. *Building and Environment* 204 (2021), 108133.

[300] Robert Spence. 2001. *Information visualization*. Vol. 1. Springer.

[301] Nitish Srivastava. 2013. Improving neural networks with dropout. *University of Toronto* 182, 566 (2013), 7.

[302] Lina Stankovic, Vladimir Stankovic, Jing Liao, and Clevo Wilson. 2016. Measuring the energy intensity of domestic activities from smart meter data. *Applied Energy* 183 (2016), 1565–1580.

[303] Chengai Sun, Qiaolin Du, Gang Tian, et al. 2016. Exploiting product related review features for fake review detection. *Mathematical Problems in Engineering* 2016 (2016).

[304] Jiao Sun, Q Vera Liao, Michael Muller, Mayank Agarwal, Stephanie Houde, Kartik Talamadupula, and Justin D Weisz. 2022. Investigating explainability of generative AI for code through scenario-based design. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*. 212–228.

[305] Dabeeruddin Syed, Haitham Abu-Rub, Ali Ghrayeb, and Shady S Refaat. 2021. Household-level energy forecasting in smart build-

ings using a novel hybrid deep learning model. *IEEE Access* 9 (2021), 33498–33511.

[306] Scott Thiebes, Sebastian Lins, and Ali Sunyaev. 2021. Trustworthy artificial intelligence. *Electronic Markets* 31, 2 (2021), 447–464. https://doi.org/10.1007/s12525-020-00441-4

[307] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine* 29, 8 (2023), 1930–1940.

[308] Chun-Hua Tsai, Yue You, Xinning Gui, Yubo Kou, and John M Carroll. 2021. Exploring and promoting diagnostic transparency and explainability in online symptom checkers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–17.

[309] Jasper van der Waa, Elisabeth Nieuwburg, Anita Cremers, and Mark Neerincx. 2021. Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial intelligence* 291 (2021), 103404.

[310] Arnaud Van Looveren and Janis Klaise. 2021. Interpretable counterfactual explanations guided by prototypes. In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part II 21*. Springer, 650–665.

[311] Nicolai Bo Vanting, Zheng Ma, and Bo Nørregaard Jørgensen. 2021. A scoping review of deep neural networks for electric load forecasting. *Energy Informatics* 4, 2 (2021), 1–13.

[312] Véronique Vasseur, Anne-Francoise Marique, and Vladimir

Udalov. 2019. A conceptual framework to understand households' energy consumption. *Energies* 12, 22 (2019), 4250.

[313] DU Vidanagama, ATP Silva, and AS Karunananda. 2022. Ontology based sentiment analysis for fake review detection. *Expert Systems with Applications* 206 (2022), 117869.

[314] Dinesh Kumar Vishwakarma, Priyanka Meel, Ashima Yadav, and Kuldeep Singh. 2023. A framework of fake news detection on web platform using ConvNet. *Social Network Analysis and Mining* 13, 1 (2023), 24.

[315] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.* 31 (2017), 841.

[316] Wei Wang, Tianzhen Hong, Ning Xu, Xiaodong Xu, Jiayu Chen, and Xiaofang Shan. 2019. Cross-source sensing data fusion for building occupancy prediction with adaptive lasso feature filtering. *Building and Environment* 162 (2019). `https://doi.org/10.1016/j.buildenv.2019.106280`

[317] Xinru Wang and Ming Yin. 2022. Effects of explanations in AI-assisted decision making: principles and comparisons. *ACM Transactions on Interactive Intelligent Systems* 12, 4 (2022), 1–36.

[318] Zhao Wang, Cuiqing Jiang, and Huimin Zhao. 2022. Know Where to Invest: Platform Risk Evaluation in Online Lending. *Information Systems Research* 33, 3 (2022), 765–783.

[319] Zhuo Wang, Hui Li, and Huiyan Wang. 2023. Vote-based integration of review spam detection algorithms. *Applied Intelligence* 53, 5 (2023), 5048–5059.

[320] Zhe Wang, Hui Zhang, Yingdong He, Maohui Luo, Ziwei Li, Tianzhen Hong, and Borong Lin. 2020. Revisiting individual and group differences in thermal comfort based on ASHRAE database. *Energy and Buildings* 219 (2020), 110017.

[321] Katharina Weitz, Dominik Schiller, Ruben Schlagowski, Tobias Huber, and Elisabeth André. 2019. " Do you trust me?" Increasing user-trust by integrating virtual agents in explainable AI interaction design. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*. 7–9.

[322] Katharina Weitz, Dominik Schiller, Ruben Schlagowski, Tobias Huber, and Elisabeth André. 2021. "Let me explain!": exploring the potential of virtual agents in explainable AI interaction design. *Journal on Multimodal User Interfaces* 15, 2 (2021), 87–98.

[323] Haibing Wu and Xiaodong Gu. 2015. Max-pooling dropout for regularization of convolutional neural networks. In *International Conference on Neural Information Processing*. Springer, 46–54. https://doi.org/10.48550/arXiv.1512.01400

[324] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564* (2023).

[325] Yeyu Wu and Bin Cao. 2022. Recognition and prediction of individual thermal comfort requirement based on local skin temperature. *Journal of Building Engineering* 49 (2022), 104025.

[326] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. 2023. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864* (2023).

[327] Jiaqing Xie, Haoyang Li, Chuting Li, Jingsi Zhang, and Maohui Luo. 2020. Review on occupant-centric thermal comfort sensing, predicting, and controlling. *Energy and Buildings* 226 (2020), 110392.

[328] Lei Xu et al. 2020. Synthesizing tabular data using conditional GAN. *Masters' Thesis* (2020).

[329] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. 2019. Modeling tabular data using conditional gan. *Advances in neural information processing systems* 32 (2019).

[330] Ke Yan, Wei Li, Zhiwei Ji, Meng Qi, and Yang Du. 2019. A hybrid LSTM neural network for energy consumption forecasting of individual households. *Ieee Access* 7 (2019), 157633–157642.

[331] Cai Yang, Mohammad Zoynul Abedin, Hongwei Zhang, Futian Weng, and Petr Hajek. 2023. An interpretable system for predicting the impact of COVID-19 government interventions on stock market sectors. *Annals of Operations Research* (2023), 1–28.

[332] Guang Yang, Qinghao Ye, and Jun Xia. 2022. Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. *Information Fusion* 77 (2022), 29–52.

[333] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems* 32 (2019).

[334] Hiroki Yoshikawa, Akira Uchiyama, and Teruo Higashino. 2021. Data balancing for thermal comfort datasets using conditional wasserstein GAN with a weighted loss function. In *Proceedings*

of the 8th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation. 264–267.

[335] Shuo Yu, Jing Ren, Shihao Li, Mehdi Naseriparsa, and Feng Xia. 2022. Graph learning for fake review detection. *Frontiers in Artificial Intelligence* 5 (2022), 922589.

[336] Hafed Zarzour, Mahmoud Al-Ayyoub, Yaser Jararweh, et al. 2021. Sentiment analysis based on deep learning methods for explainable recommendations with reviews. In *2021 12th International Conference on Information and Communication Systems (ICICS)*. IEEE, 452–456.

[337] Milan Zdravković, Ivan Ćirić, and Marko Ignjatović. 2022. Explainable heat demand forecasting for the novel control strategies of district heating systems. *Annual Reviews in Control* (2022).

[338] Gordana Zeba, Marina Dabić, Mirjana Čičak, Tugrul Daim, and Haydar Yalcin. 2021. Technology mining: Artificial intelligence in manufacturing. *Technological Forecasting and Social Change* 171 (2021), 120971.

[339] Qin Zhang, Zhiwei Guo, Yanyan Zhu, Pandi Vijayakumar, Aniello Castiglione, and Brij B Gupta. 2023. A deep learning-based fast fake news detection model for cyber-physical social services. *Pattern Recognition Letters* 168 (2023), 31–38.

[340] Wencan Zhang and Brian Y Lim. 2022. Towards relatable explainable AI with the perceptual process. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–24.

[341] Wei Zhang, Fang Liu, Yonggang Wen, and Bernard Nee. 2021. 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. 255–258.

[342] Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology* 15, 2 (2024), 1–38.

[343] Qianchuan Zhao, Yin Zhao, Fulin Wang, Jinlong Wang, Yi Jiang, and Fan Zhang. 2014. A data-driven method to describe the personalized dynamic thermal comfort in ordinary office environment: From model to application. *Building and Environment* 72 (2014). https://doi.org/10.1016/j.buildenv.2013.11.008

[344] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).

[345] Chiara Zucco, Huizhi Liang, Giuseppe Di Fatta, and Mario Cannataro. 2018. Explainable sentiment analysis with applications in medicine. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 1740–1747.