Non-Traditional Supply Chain Actors' Behavioural Interventions in Mitigating Corruption in Public Procurement

Dzaki Aulia

1. Introduction

Globally, public procurement typically accounts for between 15% of GDP in most developing economies (World Bank, 2024). In Indonesia, with a GDP of Rp6,060 trillion, public procurement spending totalled Rp1,167.8 trillion, or approximately 19.3% of GDP (LKPP, 2025). This emphasises the strategic importance of public procurement for economic and social developments. However, the magnitude and complexity of public procurement activities also followed by significant corruption risks. When public officials corruptly violate procurement processes regulations, for example awarding contracts based not on quality or price but on personal gain, the consequences can be costly, inefficient, and socially damaging (David-Barrett and Fazekas, 2015). Corruption in public procurement undermines public trust, inflates costs, reduces efficiency, and weakens service delivery (Neu et al., 2015). These risks are increased by various structural aspects of procurement processes, including high transaction volumes, substantial financial stakes, procedural complexity, close interactions between public officials and private suppliers, and the involvement of multiple stakeholders across the procurement cycle (Oecd, 2016). Furthermore, corruption is difficult to detect and prevent in public procurement because it can take many different forms and impact at all stages of the procurement process (Morgner and Chêne, 2014).

The large proportion of Indonesian public procurement spending to GDP, coupled with the country's economic status as a developing country and the recognised risks of corruption in public procurement, have positioned Indonesia as an important focus for research into corruption risk in public procurement. According to data from Indonesia's Corruption Eradication Commission (KPK), procurement-related corruption accounted for 407 documented cases between 2004 and 2024 (KPK, 2025). The trend is particularly concerning with the annual average number of cases increased from 15 between 2014 and 2018 to 68 cases in 2024 alone (KPK, 2025). To address this growing trend, corruption mitigation in public procurement has become one of the focus areas of Stranas PK, the KPK-led national strategy for corruption mitigation. Since 2019, the Stranas PK has advocated for the strengthening of the goods and services procurement system, including increasing procurement human resource capacity and modernising procurement in 2020, promoting procurement digitalisation through e-purchasing and e-payment in 2022, and improving procurement performance through supervision (e-audit) in 2024 (Setnas PK, 2025). The outcomes of these strategic efforts have converted the remaining traditional procurement process into a more transparent, effective, efficient, and inclusive electronic system. However, when it comes to annual corruption cases, there has been no major decline in procurement corruption. This demonstrates that, despite ongoing reform efforts, corruption is a persistent issue in Indonesian public procurement that must be addressed effectively and suggests that structural and technological reforms alone may be insufficient.

While institutional reforms can reshape procurement systems, it is ultimately the behaviour of procurement actors that determines whether corruption occurs. For instance, the Italian experience studied by Sargiacomo et al. (2015) demonstrates that corruption depends heavily on how bureaucrats operationalise rules and regulations, as they often possess detailed knowledge of regulatory frameworks and understand how discretion can be exercised within

them. These demonstrates how politicians, bureaucrats, or procurement officials interpret rules and respond to new regulations plays a decisive role in shaping corruption risks. Illegal behaviour in public procurement can also be understood through the fraud triangle framework, which emphasises opportunity, incentive, and rationalisation as key drivers of misconduct (Pullman et al., 2024). This can be interpreted that corruption not only emerges from system design or opportunity, but also from the behavioural responses of procurement actors based on the incentive and their rationalisation. From a DOM perspective, this means that behavioural processes, rather than system design alone, play an important role in deciding whether operational reforms have the intended outcomes in developing economies. Accordingly, effective interventions must address these behavioural dimensions, either directly through training, certification, and monitoring, or indirectly, by shaping organisational norms, expectations, and perceived constraints.

Despite this, there remains limited empirical research examining how public procurement authorities' interventions able to change the behaviour of procurement actors. To address this gap, this study systematically maps behavioural interventions implemented by Indonesia's National Public Procurement Agency (LKPP). LKPP occupies a distinctive position as a non-traditional supply chain actor: it does not directly participate in procurement transactions, yet it exerts significant influence over procurement behaviour through policies, training, certification regimes, procedural standards, and digital systems. Understanding how such interventions shape behaviour is critical for advancing corruption mitigation strategies beyond purely structural or enforcement-based approaches. This study addresses the following research question:

How do behavioural interventions implemented by a non-traditional supply chain actor influence the attitudes, norms, and perceived behavioural control of procurement officials in mitigating corruption under resource-constrained institutional environment?

To answer this question, the study employs an in-depth qualitative case study of LKPP. Data were collected through semi-structured interviews with officials from multiple directorates within LKPP, focusing on the identification, design, and implementation of behavioural interventions across policy, operational, human resource, and digital domains. Using inductive coding, the study maps these interventions and analyses how they relate to the key components of the Theory of Planned Behaviour. This research adopts Ajzen (1991) theory of planned behaviour as an analytical framework to understand how non-traditional supply chain actors influence supply chain processes and outcomes. Analysis of the behavioural aspect of these interventions requires a theory that emphasises social psychological factors that explain human behaviour, among which is Ajzen (1991) theory of planned behaviour. This theory distinguishes three types of salient beliefs: (1) Behavioural beliefs, (2) Normative beliefs and (3) Control beliefs. These three beliefs are the prevailing determinants of attitude, subjective norm, and perceived behavioural control, and these latter variables influence the individual intention that will shape their behaviour. By applying TPB, this research conceptualises procurement interventions as behaviour-shaping mechanisms that indirectly influence corruption-related behaviour by altering how procurement officials perceive integrity, social expectations, and their ability to act within procedural constraints. This research views institutional constraint as boundary conditions that systematically affect how behavioural causes are created and turned into operational outcomes. This research also expands TPB's application in a developing country setting by investigating how its fundamental constructs are developed and implemented under contexts of institutional fragility.

This study makes three contributions. First, it advances Operations Management by conceptualising corruption mitigation as a behavioural challenge and demonstrating how procurement interventions influence behaviour indirectly through attitudes, norms, and perceived behavioural control. Second, it advances the application of the Theory of Planned Behaviour by extending it from individual-level decision contexts to system-mediated interventions implemented by non-traditional supply chain actors. Third, it provides practical insights for public procurement authorities by identifying how different types of interventions may strengthen or weaken behavioural conditions associated with corruption risk, thereby informing more effective, behaviourally informed corruption mitigation strategies.

2. Theoretical Background

2.1 Corruption in Public Procurement

Public procurement refers to the acquisition of goods, services, or construction projects by government agencies. It represents a substantial proportion of public expenditure and plays a critical role in supporting government functions and national economic development. Public procurement corruption has been widely associated with institutional conditions that create opportunities for discretionary behaviour and weaken accountability. Prior research shows that low levels of competitiveness and high levels of regulatory ambiguity are strongly associated with increased perceptions of corruption in public procurement (Thomann et al., 2023).

Importantly, corruption does not necessarily arise from rule flexibility, but rather from poor rule quality, unclear responsibilities, and weakly structured public–private interactions that allow discretion to be exercised without effective oversight (Thomann et al., 2023). These conditions generate uncertainty about acceptable behaviour and reduce the perceived risks associated with corrupt actions.

Politicisation further worsen corruption risks by blurring boundaries between political authority and administrative responsibility. The involvement of politicians in bureaucratic procurement rules and practices has been consistently associated with higher corruption perceptions (Thomann et al., 2023). Politicisation enables corrupt politicians to delegate discretion to bureaucrats while retaining informal influence, thereby reducing clarity of responsibility and weakening accountability mechanisms (Loftis, 2015). As a result, procurement officials may face conflicting expectations between formal rules and informal political influences, leading to behavioural responses that differ from regulatory directives.

Corruption in public procurement is also deeply embedded in social and relational structures. Social networks involving private firms, political supporters, and influential bureaucrats have been identified as central mechanisms through which procurement corruption is organised and sustained (Lassou et al., 2023). These networks facilitate repeated interactions that normalise corrupt exchanges and reinforce shared expectations about acceptable behaviour. In such contexts, corruption may persist even when formal rules appear robust, as actors collectively interpret and enact regulations in ways that serve network interests rather than public value.

Empirical studies further reveal how corruption often manifests through the manipulation of procurement rules rather than their outright violation. Lassou (2017) documents practices such as symbolic bidding, where competition is formally simulated to comply with regulatory requirements while contracts are systematically awarded to politically connected firms or associates, often at inflated prices. These practices illustrate how procurement actors actively

interpret and exploit regulatory frameworks, demonstrating that corruption is not merely a failure of rules but a consequence of how individuals understand, rationalise, and enact those rules in practice.

These studies highlight that corruption in public procurement emerges from the interaction between institutional arrangements and individual behaviour. While regulatory structures define formal constraints, it is the beliefs, social norms, and perceived discretion of procurement actors that ultimately shape whether corruption occurs. This insight highlights the importance of complementing institutional analyses with behavioural perspectives that explain how procurement officials respond to regulatory settings.

2.2 The Organisational-Level Focused Interventions in Public Procurement

In response to corruption risks, governments and international organisations have introduced a wide range of interventions with focus on promoting procurement principles and improving compliance through digital technologies (Telgen, 2016). These interventions aimed at strengthening transparency, accountability, and control in procurement systems by modifying operational processes and encouraging the adoption of improved practices. However, corrupt procurement actors confronted with tighter regulations often adapt by identifying alternative strategies to pursue rents (Olken and Pande, 2012).

For example, the World Bank implemented large-scale reforms to update its procurement rules in November 2003, expanding e-procurement systems, introducing prior review mechanisms, and strengthening audit requirements. Despite of the successful reforms, it also found to have strong displacement effects, including the substitution of corruption techniques to less tightly controlled areas and the exploiting of control weaknesses by switching to non-competitive procedure types (Dávid-Barrett and Fazekas, 2020). These findings imply that individuals might respond by changing their behaviour to maintain opportunity for personal gain.

The digital interventions in public procurement have also become particularly prominent. Web-based platforms for reporting public transactions are now widely used in both developed and developing countries, enabling anti-corruption agencies, auditors, and citizens to monitor procurement processes and decisions more closely (Gallego et al., 2021). Empirical evidence suggests that such interventions can generate positive outcomes. Firms report paying fewer and smaller bribes in countries with more transparent procurement systems, effective complaint mechanisms, and stronger external auditing arrangements (Knack et al., 2019).

These anti-corruption interventions in public procurement have largely focused on the organisational level. Consistent with these interventions, most studies also focus more on identifying the institutional conditions under which procurement organisations adopt transparency and accountability practices, such as digital procurement systems, audit mechanisms, and formal compliance procedures (Arellano et al., 2021). In this view, corruption mitigation is primarily achieved through the redesign of organisational processes and governance structures that shape procurement decision-making.

However, the effectiveness of such organisational-level interventions varies considerably across contexts and over time. For example, analysis of misconduct cases in the United States federal procurement system identified more than 2,900 instances of supply chain misconduct over a twenty-year period involving 251 contractors, despite operating within a highly formalised regulatory environment (Gordon et al., 2021). This suggests that formal regulations

and structured procurement processes alone are insufficient to fully prevent misconduct in procurement supply chains, including corruption.

One explanation for this limitation is that organisational reforms often underestimate behavioural adaptation. Individuals operating within procurement systems are influenced not only by formal rules but also by social norms, peer behaviour, and organisational culture. Corruption can exhibit contagion effects, where observing peers engaging in corrupt behaviour increases the likelihood that others will perceive such behaviour as acceptable (Schram et al., 2022). In environments where corruption is perceived as normal or necessary for organisational survival, individuals may experience strong pressures to conform to informal practices rather than follow formal procedures (Persson et al., 2013). As a result, institutional reforms may coexist with informal behavioural patterns that continue to shape procurement decisions.

Moreover, anti-corruption programmes are often broad and implemented simultaneously with multiple other governance reforms, making it difficult to isolate the effects of specific interventions (Dávid-Barrett and Fazekas, 2020). Even when targeted interventions achieve their immediate objectives, corrupt actors may adapt strategically by shifting activities to less regulated areas or exploiting newly emerging gaps. This could be due to the inflexibility of public procurement systems as the cost of change may outweigh the benefit (Kistler et al., 2024). These dynamics highlight that procurement actors are not passive recipients of organisational reforms but active agents who interpret and respond to regulatory changes.

Taken together, prior research suggests that anti-corruption interventions in public procurement have predominantly focused on organisational-level reforms aimed at improving transparency, accountability, and procedural control. While such interventions remain necessary components of procurement governance, their effectiveness may be constrained by behavioural adaptation and informal practices within procurement environments. This limitation highlights the need to complement organisational reforms with approaches that more explicitly address the behavioural dynamics shaping procurement decision-making.

2.3 The Individual-Level Drivers of Behavioural Change

Individuals who interpret and implement process adoption are often neglected by organisationally focused interventions. For example, when achieving organisational priorities, the individual-level antecedents of procurement professionals are frequently overlooked, despite their significant role in translating those factors into actionable decisions (Kannan, 2021). The focus of prior research implicitly portrays procurement officials and managers as rational actors who respond mechanically to institutional pressures by adopting prescribed practices (Donohue et al., 2020). This perspective simplifies the complex decision-making processes that shape how individuals respond to reforms in practice.

Individuals play an important role in achieving organisational goals, as noted by Hinterhuber and Liozu (2017), "organisations do not act—individuals do." As a result, individuals' ability to translate organisational priorities into operational activities is critical. Because organisations rely on individuals to implement strategies, understanding micro-level behaviours is essential for explaining macro-level outcomes (Hinterhuber and Khan, 2025). These micro-level behaviours, which include individuals' attitudes, motivations, and perceptions, collectively influence organisational outcomes (Zhao and He, 2022). Micro-foundations are also gaining significant interest in operations and supply chain management.

Recent studies argue practice adoption is mediated by individuals' interpretations of organisational practices. Managers do not passively implement practices. Instead, they evaluate the pressuring factors based on their own beliefs, experiences, and perceptions (Jacqueminet, 2020). This is critical for the organisation because managers' biases heavily influence decision-making processes (Malmendier et al., 2023). Insufficient knowledge and distorted perceptions can have a significant impact on the organisation. Organisational reforms do not directly cause behavioural change. Instead, they influence behaviour through cognitive processes that allow people to interpret the value and consequences of implementing new practices.

Individual behavioural frameworks are increasingly important in procurement research. For instance, Constant and Johnsen (2024) examine procurement's role in driving organisational innovation using the awareness–motivation–capability framework, which similarly emphasises individual-level decision-making as a foundation for organisational outcomes. Procurement officials have strategic roles to its organisation because they have a discretion to make decisions which often involve ethical judgement, internal or external pressures, and their interactions with stakeholders, making them rely more to their own attitudes based on the social norms and how they perceived their controls.

Understanding individual decision-making processes requires theoretical framework that explicitly account for individual behaviours. One of the most prominent framework is the theory of planned behaviour that provides measurable constructs of attitudes, norms, and perceived controls to predict intentions that results in specific behaviour (Ajzen, 1991). For example, Arellano et al. (2021) interviews operations managers to examine the antecedents of adopting new operations management practices, highlighting the role of individual beliefs and intentions. Understanding the behavioural drivers is essential to explain why similar interventions applied in different organisation might result in different outcomes.

In another case, Foerstl et al. (2021) analyse 145 production insourcing decisions and find that individual attitudes and social pressures significantly shape intentions, which in turn lead to actual organisational behaviour. Similarly, Shou et al. (2022) focus on the normative component of TPB and demonstrate that subjective norms influence the adoption of green innovations. The TPB provides a robust framework for explaining why professionals engage in particular behaviours (Hinterhuber and Khan, 2025). TPB therefore provides a bridge between individual cognition and collective outcomes, enabling analysis of how behavioural interventions targeted at individuals' beliefs can influence organisational-level behaviour.

In public procurement, organisational integrity, compliance, and corruption risk arise as a result of repeated individual decisions made by procurement officials working within institutional settings. It is also influenced by external factors such as other procurement actors involved in the decision-making process. Corruption in public procurement cannot be fully understood without considering procurement actors' individual beliefs and intentions. This research identified individual-level drivers as an important aspect of behavioural change targeted by organisational-level interventions that began with the formation of individual beliefs.

2.4 Beliefs Formation in Procurement Governance

To examine how procurement interventions influence corruption-related behaviour in public procurement, this study draws on the Theory of Planned Behaviour (TPB). This theory proposes that behaviour is preceded by behavioural intentions which are shaped by the attitudes, subjective norm, and perceived behavioural control (Ajzen, 1991). Attitudes explain

individual's evaluation of performing a behaviour, subjective norms reflect perceived social pressures from important referent groups, and perceived behavioural control represents an individual's perception of their ability to perform or avoid a given behaviour (Ajzen, 1991).

While this theory was initially conceptualised at the individual level, recent research has extended its application to organisational contexts by using managerial perceptions and decision-making as contributing factors for organisational behaviour (Foerstl et al., 2021). TPB offers a theoretical lens to investigate how individual-level decisions, when aggregated across actors and over time, shape organisational practices and performance (Hinterhuber and Khan, 2025). Additionally, this theory also has increasingly been recognised in operations and supply chain management research (Arellano et al., 2021; Shou et al., 2022).

Addressing corruption in public procurement requires behavioural change at both the organisational and individual levels. Building on this perspective, the objective of this research is to identify and characterise the key beliefs that emerge from procurement actors' experiences with procurement interventions that drive practice adoption in corruption mitigation. Attitudes toward corruption risks, perceived social expectations regarding procurement behaviour, and perceptions of behavioural control within procurement processes all contribute to the formation of procurement governance beliefs.

In public procurement, beliefs that shape procurement actors' behaviour are likely to be influenced by procurement governance enforced through digital platforms. The principles of transparency, accountability, and fairness in public procurement might influence procurement actors' beliefs about the practices they must follow, as well as the risks, expectations, and feasibility associated with their actions. Procurement interventions can have an indirect impact on behavioural outcomes by reshaping individuals' beliefs that shape their attitudes, subjective norms, and perceived behavioural control.

The three factors of TPB grounded in salient beliefs that individuals develop through their experiences and interpretations of specific environments (Ajzen, 1991). Behavioural beliefs shape attitudes toward behaviour, which reflect an individual's assessment of engaging in the behaviour in question (Ajzen, 1991). Procurement actors may question the benefit to themselves that led to their decision to adopt a practice. The efficacy of the procurement intervention will be the primary determinant of whether it has the potential to mitigate corruption or is simply an administrative change that must be adhered to.

Normative beliefs give rise to subjective norms, which reflect an individual's perception of the social pressure to engage in a specific behaviour (Ajzen, 1991). The desire to conform with the expectations of others conforms with the legitimacy driver, which explain the adoption of practices intended to increase their credibility in the eyes of others (Voss, 2005). If meeting procurement intervention objectives will improve their position among others, procurement actors will be positive about the intervention. The resources provided for the interventions also determine whether the procurement actors comply with the intervention's obligations.

Control beliefs influence perceived behavioural control and represent an individual's ability to perform the behaviour (Ajzen, 1991). These beliefs consider whether it is easy or difficult to adopt a practice and define individuals' commitment to achieving a specific outcome (Jimmieson et al., 2008). Procurement actors will assess the impact of their actions on interventions and ensure that the results have an impact in obtaining desired results. If they

believe their actions are insignificant, they will be less committed and will have a lower level of support because of their low sense of control (Vardaman et al., 2012).

TPB explains how beliefs shape attitudes, subjective norms, and perceptions of behavioural control, which influence intentions that lead to behaviour in professional settings where decision-making is influenced by both individual cognition and organisational context (Hinterhuber and Khan, 2025). Using TPB as a theoretical framework allows for the identification of belief constructs and the analysis of the change in behavioural, normative, and control beliefs as a result of interventions, whether positive or negative (Steinmetz et al., 2016). The focus on belief constructs will explain how interventions can shape intentions and behaviour of procurement actors in the effort to mitigate corruption.

3. Research Design

This research examines how non-traditional supply chain actors intervene in the public procurement process to influence the behaviour of procurement actors to mitigate corruption. As corruption in public procurement has gained sustained global attention, States have collectively committed to combating it through the United Nations Convention against Corruption (UNCAC), the most comprehensive international legal instrument on anti-corruption. All States Parties are obliged to adopt preventive measures, enhance transparency, and strengthen control mechanisms, with public procurement identified as a particularly high-risk area. At the 11th Conference of the States Parties (CoSP11), held in Doha, Qatar, civil society organisations urged States to make more effective use of UNCAC provisions at the national level, specifically by strengthening integrity, transparency, and accountability in public procurement systems (UNCAC Coalition, 2025). However, despite these formal commitments, many States have struggled to translate UNCAC provisions into practice.

Our research design consists of an in-depth case study. Our primary data come from qualitative interviews, complemented by publicly available documents and internal secondary data. Of particular importance to this study is the role of interventions in influencing procurement actors' behaviour and mitigating corruption risks. Accordingly, we selected a case in which the National Public Procurement Agency implements multiple interventions across Procurement Service Units in different ministries, agencies, and regional governments, forming embedded relationships that afford multiple levels of analysis. This case is considered particularly suitable because there is limited empirical research examining how non-traditional supply chain actors intervene in public procurement processes to mitigate corruption.

The case study approach is an effective research tool that allows for objective, in-depth analysis of contemporary events and draws on multiple sources of data to improve reliability and validity (Mccutcheon and Meredith, 1993). Given the variety of interventions implemented by LKPP, this research uses the case to support theory building by examining the mechanisms that link interventions, behavioural change, and corruption prevention. Theory building can be conducted by an in-depth examination of the relationship among concepts (Ketokivi and Choi, 2014). In this case, LKPP is uniquely positioned as a non-traditional supply chain actor operating within a dispersed public procurement supply network. As a result, the case study can give valuable insights that enhance and expand previous studies on corruption, governance, and supply chain intervention.

3.1 The Case Study

This research was conducted in Indonesia, the largest economy in Southeast Asia, where the scope of government activities demands public procurement as a critical governance function. As a developing country with a large geographical area, Indonesia's institutional quality, administrative capacity, and monitoring mechanisms differ significantly between areas. This background has contributed to persistently significant corruption risks in public procurement. As a commitment of anti-corruption efforts, Indonesia remains formally committed to the full implementation of the UNCAC through its three-pronged corruption eradication strategy of prevention, enforcement, and education. In practice, a National Corruption Prevention Strategy with a focus on licensing and trade, state finances, law enforcement, and bureaucratic reform is in place. Corruption prevention in public procurement includes in the second focus of state finances with a strong emphasis on digitising procurement governance.

Indonesia's public procurement system operates within a decentralised budgeting and procurement structure, in which procurement authority is distributed across regions. The National Public Procurement Agency of Indonesia or LKPP serves as the national policy-making body and digital system developer. LKPP develops regulatory frameworks and digital infrastructure for 94 Procurement Services Units (PSUs) in the central government and 546 PSUs in regional governments, operating under varying local conditions. Procurement maturity differs substantially across regions, reflecting differences in human resource capacity, political pressures, and organisational experience. Many PSUs also face practical constraints, including limited staffing and administrative burdens. These conditions shape how centrally designed anti-corruption initiatives are interpreted and implemented at the local level, whether they are mandated through the National Corruption Prevention Strategy led by the Corruption Eradication Commission (KPK) or introduced directly by LKPP.

The case study focuses on LKPP, the national authority responsible for public procurement policy and digital procurement system development in Indonesia, and 30 representative Procurement Services Units (PSUs), the organisational units responsible for day-to-day procurement implementation at the operational level. Given the interdependencies between these actors, the case is conceptualised as a series of embedded interactions between LKPP and PSU. In this configuration, LKPP designs and deploys interventions, while PSUs translate them into routine procurement practices, enabling analysis of how centrally designed interventions produce behavioural change on operational level.

Indonesia represents an important case for Development Operations Management research because it combines large-scale public procurement operations, institutional constraints common to emerging economies, and a highly integrated national procurement authority. Anti-corruption solutions developed in advanced economies may not translate directly to such contexts, where challenges such as inadequate infrastructure, fragile institutions, political instability, currency volatility, and the prevalence of informal sectors remain salient. This case study therefore offers a valuable opportunity to examine how corruption control and behavioural change are pursued not only through professional norms and enforcement mechanisms, but also through procurement process design, digital infrastructure, and advisory interventions.

3.1.1 LKPP or The National Public Procurement Agency of Indonesia. LKPP is the national authority responsible for public procurement policy and system development in Indonesia. Its mandate includes formulating procurement regulations and standard operating procedures, developing competency frameworks and capacity-building policies for procurement personnel, providing technical assistance, advocacy, and legal opinions related to public procurement, designing and managing national electronic procurement systems. LKPP has developed a suite of electronic systems for reporting procurement plans, automating procurement processes, enabling purchasing via e-Catalogs, and integrating procurement with payment systems. Through these functions, LKPP plays a central role in shaping the institutional environment of public procurement, strengthening monitoring, control, standardisation, and procurement capability across a diverse range of public institutions nationwide. The scope and reach of LKPP's mandate make it a particularly suitable focal organisation for this research.

As a central actor with authority over public procurement policy and system, LKPP occupies a unique position within Indonesia's public procurement supply network, enabling it to intervene across multiple stages of the procurement process and influence actor behaviour at scale. In response to corruption risks, LKPP has increasingly focused on preventive and behavioural interventions aimed at influencing how procurement actors interpret and enact procurement rules in practice. LKPP designs interventions such as policy instruments, digital systems, guidance frameworks, training programmes, and monitoring mechanisms intended to promote integrity, transparency, and accountability across procurement processes. These interventions are disseminated nationwide and are expected to be adopted by PSUs operating within diverse organisational and regulatory contexts. This research seeks to understand how such interventions shape the behaviour of public procurement actors as a mechanism for corruption prevention.

3.1.2 Procurement Service Units (PSUs). Procurement Service Units (PSUs) are organisational units embedded within ministries, agencies, and regional governments that function as centres of excellence for public procurement. PSUs are responsible for conducting procurement activities in accordance with national regulations, managing tendering processes, supporting user departments, and ensuring compliance with procurement policies and standards. In practice, PSUs operate under varying institutional conditions, including differences in organisational capacity, resource availability, leadership support, and corruption risk exposure. As a result, while PSUs are formally required to implement LKPP's interventions, the way these interventions are interpreted, enacted, and embedded into procurement routines may differ across contexts.

This variation makes PSUs a critical empirical site for examining how centrally designed anti-corruption interventions translate into behavioural change at the operational level. Within this research, PSUs represent the recipient side of LKPP's interventions, providing insight into how procurement actors respond to, adapt, or resist efforts aimed at promoting integrity and preventing corruption. Examining multiple PSUs allows for comparative analysis across embedded dyads, enhancing understanding of the mechanisms through which interventions influence behaviour in public procurement settings. The total of 30 representative Procurement Services Units (PSUs) are involved in this research to examine the 5 priority interventions of LKPP in mitigating corruption in public procurement.

3.2 Data Collection

Our primary data collection method was semi-structured interviews that split into two phases; first phase interviews with relevant LKPP officials, second phase interviews with 30 officials from different PSUs. Table 1 lists the profiles of the interviewees. The first phase interviews mapped the LKPP's interventions designed to influence the behaviour of procurement officials to mitigate corruption while the second phase interviews gathered how LKPP's interventions are perceived and experienced in practice, and how they influence attitudes, subjective norms, and perceived behavioural control in daily procurement decisions. This data was complemented by secondary data, including LKPP reports, regulations, training materials, evaluation documents, and website information. The first phase interviews with LKPP were used to introduce the study, clarify expectations and understand the backgrounds, objectives, mechanisms, and intended behavioural outcomes of each intervention. The 30 interviews were conducted with representatives from various PSUs, covering roles from senior-level to middle-level officials responsible for procurement activities in each PSUs. Furthermore, a focus group discussion was held with LKPP to review the interventions and decide five top priority interventions to be discussed with PSUs in the second phase interviews. This was followed by an online workshop at the conclusion of the study to present the findings and ensure credibility through participant feedback. In total, the interviews contributed approximately 60 h of data. The interview data was complemented by data provided by LKPP, including their reports and internal studies. These documents were analysed for triangulation with the facts and figures mentioned by the interviewees. Data collection commenced in November 2025 and lasted for approximately 6 months.

3.3 Data Analysis

All the interviews were transcribed and coded using the qualitative data analysis software MAXQDA to ensure an audit trail of transcripts and codes. A line-by-line analysis of the transcripts allowed us to extract participant insights related to the research question. These were organised into first-order codes. We stayed close to the data by focusing on participants' understandings of the situation. Typically, the participants spoke about the challenges in implementing the interventions and how corruption risk affected their procurement processes.

These first-order codes enabled us to comprehend the situation, but they did not fully explain how interventions effectively mitigate corruption in procurement processes and change the behaviour of procurement actors. Based on the first-order codes, we wrote case descriptions to gain an initial understanding of LKPP viewpoints and PSUs viewpoints. We also presented the initial analysis of the first-order codes to LKPP's senior and middle level officials during a workshop to ensure that the correct ideas had been captured. This allowed the firm to provide updates following the interviews, which strengthened data reliability and relevance.

4. Observed Change in Beliefs

The current Presidential Regulation requires the use of the Electronic Procurement System's transactional features in all supplier selection methods, including direct procurement. This reform effectively eliminates transactions outside the system by mandating digital traceability across procurement methods that have traditionally been associated with greater discretion and less scrutiny. This intervention reconfigures the process architecture by integrating transactional visibility, audit trails, and standardised workflows into routine procurement activities. It shifts from procedural compliance to process enforcement via system design, in line with the national anti-corruption strategy. This intervention might change the belief configuration of procurement actors, and this section described the observed change in belief based on interviews with several procurement managers, using their perspectives to assess what has changed in practice.

4.1 Behavioural Beliefs of Transparency and Traceability

Behavioural beliefs reflect how people assess the expected consequences of a behaviour (Ajzen, 1991). With a digital system in place, including a transactional feature for direct procurement, it has the potential to improve transparency and traceability of procurement activities while reducing opportunities for manipulation. As one of the procurement managers stated, *"With the use of the transactional feature, the parties involved should be more careful because everything is recorded, both in terms of documents and the selection process. I think each party must be careful in every process carried out."*

Procurement managers stated that digital procurement activities create a transparent timeline of events, reducing the possibility of changing procurement documents. Documents in manual processes could be changed during the process, whereas electronic systems record every step of the procurement lifecycle. As one respondent stated, *"When direct procurement was manual, it was somewhat difficult to measure data consistency and accountability. Because documents could easily be changed. Now by using the system, at least it must be measurable from the beginning, from planning until the completion of the work."*

Procurement managers also noticed behavioural changes among suppliers. In contrast to the manual process, where communication prior to contract is less transparent, suppliers now recognise that informal attempts to influence procurement decisions are ineffective. The visibility of procurement data enables authorities to monitor which suppliers receive contracts. As one procurement manager stated, *"When it is recorded the timeline is visible. Also, who submitted the offer and was selected as the provider in the direct procurement is visible. And the contract value at what price is visible. It is more open."*

This digital traceability was also believed to increase accountability in procurement processes. Procurement managers explained how the system requires procurement planning, documentation, and implementation to be coordinated through a digital platform. Procurement actors believe the system makes manipulation more difficult, as explained by one procurement manager, *"There must be consistency from the beginning of planning, preparation of the RUP until the completion of the work. Since all data has been uploaded into the application, it is perhaps more accountable, and somewhat more difficult to manipulate."*

This increased visibility adds to the perception that corruption risks have decreased since the intervention. Several procurement managers explicitly linked the implementation of the

electronic system to a decrease in opportunities for corrupt practices. According to one respondent, *"In terms of corruption risk, it definitely decreases with this transactional application. The obstacle is reduced in terms of corruption risk."* These perceived outcomes of the digital system encourage procurement actors to be more cautious and adhere to formal procurement procedures, thereby reducing opportunities for corruption in public procurement.

4.2 Normative Beliefs of Institutional Expectations and Professional Norms

Normative beliefs reflect others' perceived expectations and pressures to behave appropriately (Ajzen, 1991). These expectations might derive from a variety of sources, including regulations, professional norms among procurement officials, organisational leadership, and oversight institutions. Regulatory requirements emerged as one of the most powerful normative drivers influencing procurement behaviour following the intervention. Resistance was observed during the initial period when the regulation was implemented, which required each procurement work unit to adopt the system and transition from manual processes.

According to one procurement manager, *"When we first used the transactional feature, there was resistance from several OPDs. Why must we use the transactional feature when it can still be done manually? But because we follow the Presidential Instruction, we required it to be carried out using the transactional feature. If done manually, we could not continue the process."* This condition explains the role of regulation in the adoption of the transactional procurement system, which is motivated by government policy rather than internal organisational initiatives.

Beyond regulation, professional norms were reinforced through oversight mechanisms. This encourages procurement officials to ensure that documentation and implementation are consistent. As one manager explained, *"From the internal supervisory side, their audit process now prioritises the conformity of data in the application with the implementation process carried out in the field."* Previous cases have also shaped collective attitudes toward risk-taking behaviour, as explained *"Manipulation has decreased. Because some have faced cases, colleagues intervene saying they do not want to end up like others."*

Normative pressures were also evident regarding the importance of personal integrity and professional responsibility in ensuring compliance with procurement regulations. As stated by one of the managers, *"I think besides regulatory demands there must also be personal willingness to comply."* There is also a shift in mindset toward prioritising long-term career goals over short-term personal gain. According to one procurement manager, *"Everyone has become aware. Beyond the system, we want career sustainability. Colleagues do not need something momentary, we have careers, families and must consider that."*

These institutional expectations also have an impact on supplier behaviour, as they are less likely to approach procurement officials informally. As one procurement manager explained, *"Providers have also become aware that using money does not always smooth the process or ensure winning. As procurement actors' attitudes change, vendors are reluctant to approach."* The intervention has strengthened normative beliefs by reinforcing institutional expectations and professional norms governing procurement behaviour, which shapes procurement actors' perceptions of appropriate conduct and discouraging corrupt practices.

4.3 Control Beliefs of the Procurement System

Control beliefs shape perceived behavioural control and concern the presence of factors that may facilitate or impede the execution of a specific behaviour (Ajzen, 1991). This belief reflects procurement actors' perceptions of their ability to engage in the behaviour required by the intervention. The procurement managers believe the electronic procurement system improves procedural control by structuring procurement processes and limiting opportunities for discretionary manipulation. It enforces procedural discipline because procurement activities must adhere to predefined stages and timelines on the platform.

The system prevents changes to procurement documentation once submission deadlines have passed, as stated by a procurement manager, *"If using the transactional feature in the system, documents that have been uploaded and the upload period has ended cannot be changed."* The digital system also helps procurement actors monitor procurement timelines and activities by automatically recording procurement events, making it easier for officials to track and verify. One of the managers explained, *"Who submitted the offer and was selected as the provider in the direct procurement [using transactional feature] is visible."*

However, several institutional constraints limit the intervention's effectiveness. First, there is a lack of integration between procurement systems and financial payment mechanisms. As a purchasing manager clarified, *"The transactional use of procurement systems is not directly linked to the payment mechanism, so regional apparatuses can still carry out procurement without using the transactional feature."* Because payment for completed work can still be processed through the financial administration system, procurement actors may still avoid the transactional feature of procurement systems.

The second constraint that procurement actors face at the organisational level is that, in some cases, project deadlines or urgent work requirements encourage organisations to bypass the electronic procurement system. For example, some regional units prefer manual procurement because it provides more flexibility in adjusting documentation timelines. One manager described this situation as *"Regional Apparatus Organisations argue that the work is urgent and must be carried out immediately, and if done through the system it will not meet the deadline. With this reason it is eventually done manually because manually it can be backdated."*

Despite the constraints, this intervention provides additional justification for procurement officials to follow formal procedures and ignore informal requests or pressure from others. As one manager explained, *"This intervention is more useful for procurement officials, because they will not want to process if it is not through the system."* The intervention influences control beliefs by strengthening procedural control through digital governance mechanisms, while also demonstrating institutional constraints that limit the system's overall effectiveness. This influences procurement actors' perceptions of their ability to comply with system-based procurement practices, as well as the intervention's effectiveness in reducing corruption risks in public procurement.

4.4 Additional Belief on Governance Responsiveness

Our data analysis reveals how procurement actors' behaviour is influenced by their perceptions of how institutions react to procurement practices. These responses can include monitoring, continuous evaluation, and corrective intervention. The monitoring capability enables

oversight bodies to identify potential irregularities and report them to supervisory institutions. As one manager described, *"Our prevention is only sending notification. We cooperate with the inspectorate that there is an indication of one provider winning many packages across several regional apparatus organisations or concentrated in certain regional apparatus organisations."*

If these potential issues when unusual patterns emerge are not communicated to internal oversight apparatuses, the same incident may occur again. The system can do this monitoring, as explained by a procurement manager, *"We can see indications from providers receiving those contracts. We can only see that. Since it is electronic, it can be tracked electronically where a provider gets work."* This monitored transaction is used by management to evaluate procurement decisions made by procurement officials and intervene when irregularities are discovered. This type of institutional responsiveness increases accountability in procurement processes.

As explained by one of the procurement managers, *"Management evaluates selection results conducted by procurement officials in the Procurement Work Unit. This is done on a random sample basis. If results do not comply with regulations, assignments may be withheld and further training provided."* To respond to the monitoring and evaluation, procurement managers emphasised the importance of providing advisory and corrective guidance to procurement units to prevent irregular procurement practices. As stated here, *"In regions, our intervention is mostly advisory, asking them not to manipulate and to ensure transactions are transparent and uploaded in the system."*

These governance interactions contribute to larger learning processes. Procurement managers explained that the increased visibility of procurement data enables institutions to identify systemic risks and gradually strengthen preventative measures. Over time, this process helps to reduce risk and increase transparency in procurement practices. As one respondent observed, *"Work units are now open and coordinate immediately with the Ministry when issues arise. It is more open than before. There is improvement and increased risk mitigation in procurement. Risks have decreased year by year; the trend is downward."*

Based to these findings, there is an additional belief that cannot be fully explained by the behavioural, normative, or control beliefs, particularly whether the governance mechanism actively monitors and detects irregularities, evaluates and takes corrective action when procurement processes deviate. This additional belief can be classified as governance responsiveness beliefs, as it reflects procurement actors' perceptions that procurement behaviour is embedded in a responsive governance environment that actively implements monitoring, feedback, and corrective interventions.

References

Ajzen, I. (1991). 'The theory of planned behavior' *Organizational Behavior and Human Decision Processes*, 50 (2), pp. 179-211 https://doi.org/10.1016/0749-5978(91)90020-T.

Arellano, M. C., Meuer, J. and Netland, T. H. (2021). 'Commitment follows beliefs: A configurational perspective on operations managers' commitment to practice adoption' *Journal of Operations Management*, 67 (4), pp. 450-475 https://doi.org/10.1002/joom.1130.

Bank, W. (2024). *Making Procurement Work Better – An Evaluation of the World Bank's Procurement System*: World Bank. Available at: http://hdl.handle.net/10986/42507.

Coalition, U. (2025). *Anti-corruption Priorities from the Global Civil Societ Coalition for the UNCAC*: Global Civil Society Coalition for the UNCAC Submission to CoSP11.

Constant, F. and Johnsen, T. (2024). 'Purchasing contribution to innovation exploration: awareness, motivations and capabilities' *International Journal of Operations & Production Management*, 45 (2), pp. 493-516 https://doi.org/10.1108/ijopm-10-2023-0849.

David-Barrett, E. and Fazekas, M. (2015). *Corruption Risks in UK Public Procurement and New Anti-Corruption Tools*. Budapest: Government Transparency Institute.

Dávid-Barrett, E. and Fazekas, M. (2020). 'Anti-corruption in aid-funded procurement: Is corruption reduced or merely displaced?' *World Development*, 132 https://doi.org/10.1016/j.worlddev.2020.105000.

Donohue, K., Özer, Ö. and Zheng, Y. (2020). 'Behavioral Operations: Past, Present, and Future' *Manufacturing & Service Operations Management (M&SOM) (INFORMS)*, 22 (1), pp. 191-202 https://doi.org/10.1287/msom.2019.0828.

Foerstl, K., Franke, H. and Cataldo, Z. (2021). 'What drives managers to insource production? Evidence from a behavioural experiment' *Journal of Purchasing and Supply Management*, 27 (4), p. 100715 https://doi.org/https://doi.org/10.1016/j.pursup.2021.100715.

Gallego, J., Rivero, G. and Martínez, J. (2021). 'Preventing rather than punishing: An early warning model of malfeasance in public procurement' *International Journal of Forecasting*, 37 (1), pp. 360-377 https://doi.org/10.1016/j.ijforecast.2020.06.006.

Gordon, N., Oliveira, F. and Watson, Z. (2021). 'COVID and Contractor Misconduct'.

Hinterhuber, A. and Khan, O. (2025). 'What drives sustainable procurement? Insights from the theory of planned behavior' *International Journal of Operations & Production Management*, 45 (13), pp. 28-52 https://doi.org/https://doi.org/10.1108/IJOPM-02-2024-0164.

Hinterhuber, A. and Liozu, S. M. (2017). 'The micro-foundations of pricing' *Journal of Business Research*, 76 pp. 159-162 https://doi.org/10.1016/j.jbusres.2016.11.018.

Jacqueminet, A. (2020). 'Practice implementation within a multidivisional firm: The role of institutional pressures and value consistency' *Organization Science*, 31 (1), pp. 182-199.

Jimmieson, N. L., Peach, M. and White, K. M. (2008). 'Utilizing the Theory of Planned Behavior to Inform Change Management:An Investigation of Employee Intentions to Support Organizational Change' *The Journal of Applied Behavioral Science*, 44 (2), pp. 237-262 https://doi.org/10.1177/0021886307312773.

Kannan, D. (2021). 'Sustainable procurement drivers for extended multi-tier context: A multi-theoretical perspective in the Danish supply chain' *Transportation Research Part E: Logistics and Transportation Review*, 146 p. 102092 https://doi.org/https://doi.org/10.1016/j.tre.2020.102092.

Ketokivi, M. and Choi, T. (2014). 'Renaissance of case research as a scientific method' *Journal of Operations Management*, 32 (5), pp. 232-240 https://doi.org/https://doi.org/10.1016/j.jom.2014.03.004.

Kistler, J. T., et al. (2024). 'Does history really repeat itself? An empirical investigation of recurring misconduct violations in public procurement' *Journal of Purchasing and Supply Management*, 30 (1), https://doi.org/10.1016/j.pursup.2023.100893.

Knack, S., Biletska, N. and Kacker, K. (2019). 'Deterring Kickbacks and Encouraging Entry in Public Procurement Markets: Evidence from Firm Surveys in 90 Developing Countries' *World Bank Economic Review*, 33 (2), pp. 287-309 https://doi.org/10.1093/wber/lhy016.

Lassou, P. J. C. (2017). 'State of government accounting in Ghana and Benin: a "tentative" account' *Journal of Accounting in Emerging Economies*, 7 (4), pp. 486-506 https://doi.org/10.1108/jaee-11-2016-0101.

Lassou, P. J. C., et al. (2023). 'Monetization of politics and public procurement in Ghana' *Accounting, Auditing and Accountability Journal*, https://doi.org/10.1108/AAAJ-07-2021-5341.

Loftis, M. W. (2015). 'Deliberate Indiscretion? How Political Corruption Encourages Discretionary Policy Making' *Comparative Political Studies*, 48 (6), pp. 728-758 https://doi.org/10.1177/0010414014556046.

Malmendier, U., Pezone, V. and Zheng, H. (2023). 'Managerial Duties and Managerial Biases' *Management Science*, 69 (6), pp. 3174-3201 https://doi.org/10.1287/mnsc.2022.4467.

Mccutcheon, D. M. and Meredith, J. R. (1993). 'Conducting case study research in operations management' *Journal of Operations Management*, 11 (3), pp. 239-256 https://doi.org/https://doi.org/10.1016/0272-6963(93)90002-7.

Morgner, M. and Chêne, M. (2014). *Public Procurement Topic Guide Compiled by The Anti-Corruption Helpdesk*: Transparency International.

Neu, D., Everett, J. and Rahaman, A. S. (2015). 'Preventing corruption within government procurement: Constructing the disciplined and ethical subject' *Critical Perspectives on Accounting*, 28 pp. 49-61 https://doi.org/10.1016/j.cpa.2014.03.012.

Oecd (2016). *Preventing Corruption in Public Procurement*. Paris: OECD.

Olken, B. A. and Pande, R. (2012). 'Corruption in developing countries' *Annu. Rev. Econ.*, 4 (1), pp. 479-509.

Persson, A., Rothstein, B. and Teorell, J. (2013). 'Why Anticorruption Reforms Fail—Systemic Corruption as a Collective Action Problem' *Governance*, 26 (3), pp. 449-471 https://doi.org/https://doi.org/10.1111/j.1468-0491.2012.01604.x.

Pullman, M., Mccarthy, L. and Mena, C. (2024). 'Breaking bad: how can supply chain management better address illegal supply chains?' *International Journal of Operations & Production Management*, 44 (1), pp. 298-314 https://doi.org/10.1108/ijopm-02-2023-0079.

Sargiacomo, M., et al. (2015). 'Accounting and the fight against corruption in Italian government procurement: A longitudinal critical analysis (1992-2014)' *Critical Perspectives on Accounting*, 28 pp. 89-96 https://doi.org/10.1016/j.cpa.2015.01.006.

Schram, A., Zheng, J. D. and Zhuravleva, T. (2022). 'Corruption: A cross-country comparison of contagion and conformism' *Journal of Economic Behavior & Organization*, 193 pp. 497-518 https://doi.org/https://doi.org/10.1016/j.jebo.2021.11.017.

Shou, Y., et al. (2022). 'Actions speak louder than words? The impact of subjective norms in the supply chain on green innovation' *International Journal of Operations & Production Management*, 43 (6), pp. 879-898 https://doi.org/10.1108/ijopm-04-2022-0265.

Steinmetz, H., et al. (2016). 'How Effective are Behavior Change Interventions Based on the Theory of Planned Behavior?' *Zeitschrift für Psychologie*, 224 pp. 216-233 https://doi.org/10.1027/2151-2604/a000255.

Telgen, J., Van Der Krift, J., and Wake, A (2016). *Public Procurement Reform: Assessing Interventions Aimed at Improving Transparency.* London: Department for International Development.

Thomann, E., Marconi, F. and Zhelyazkova, A. (2023). 'Did pandemic responses trigger corruption in public procurement? Comparing Italy and Germany' *Journal of European Public Policy*, p. 30 https://doi.org/10.1080/13501763.2023.2241879.

Vardaman, J. M., et al. (2012). 'Interpreting change as controllable: The role of network centrality and self-efficacy' *Human Relations*, 65 (7), pp. 835-859 https://doi.org/10.1177/0018726712441642.

Voss, C. A. (2005). 'Paradigms of manufacturing strategy re-visited' *International Journal of Operations & Production Management*, 25 (12), pp. 1223-1227 https://doi.org/10.1108/01443570510633620.

Zhao, L. and He, Q. (2022). 'Explicating the microfoundation of SME pro-environmental operations: the role of top managers' *International Journal of Operations & Production Management*, 42 (4), pp. 500-525 https://doi.org/10.1108/ijopm-09-2021-0590.

An Interpretable Machine Learning Approach for Forecasting Extreme Price Occurrences in the Day-Ahead Electricity Market

Jingchuan Ma

10974220

Supervisors: Prof. Yu-wang Chen, Dr. Fanlin Meng

Alliance Manchester Business School
The University of Manchester

March 27, 2026

Abstract

In recent years, the frequency and magnitude of extreme electricity prices have increased due to the rising share of renewable energy and external market dynamics, exposing market participants to significant price risks. Accurately forecasting extreme price occurrences is essential for enhancing market stability and supporting informed decision-making in price risk management. To address this challenge, this paper introduces an adaptive, dynamic weighted threshold (DWT) method for identifying extreme prices under varying market conditions. Furthermore, it proposes a weighted-XGBoost (W-XGB) classification model to forecast extreme price occurrences in the context of imbalanced data. Comparative analysis across various experiments demonstrates the proposed method's superior forecasting performance and stronger discriminatory power compared to other baseline models. To enhance the model's interpretability, SHAP (SHapley Additive exPlanations) is applied to analyze the relative contributions of different features. The analytical results reveal that extremely high prices are influenced by multiple interrelated factors, including supply-demand conditions, fossil fuel price volatility, and historical market behaviours, making their forecast more complex. In contrast, extremely low prices are predominantly driven by forecasted residual load, indicating a more deterministic relationship with supply-demand conditions. Additionally, while geopolitical risk is considered, it exhibits minimal direct impact on extreme price occurrences. By incorporating SHAP-based interpretability analysis, this study provides a deeper understanding of extreme electricity price dynamics. The proposed solution is adaptable to various electricity markets, offering valuable insights for market operators and participants seeking to enhance risk management strategies and improve forecasting accuracy in volatile electricity markets.

1 | INTRODUCTION

With the deregulation of European electricity markets in the late 20th century, market participants became increasingly exposed to significant price risks. Unlike traditional commodities, electricity cannot be efficiently stored, and its transmission is constrained by the physical limitations of the grid (Souhir et al. 2019). Consequently, electricity spot prices exhibit higher volatility than other commodities and are characterized by extreme price behaviours, including sharp price spikes and even negative prices.

Furthermore, the growing adoption of renewable energy resources in recent years has caused electricity price volatility in Europe to surge, and the frequency of extreme price events has also increased significantly. More specifically, renewable energy resources such as wind power and photovoltaic power are intermittent because their production depends on uncertain weather conditions. This leads to an imbalance between electricity supply and demand, resulting in price volatility and the occurrence of extreme prices (Ketterer 2014, Maniatis and Milonas 2022). Furthermore, geopolitical events have a significant impact on the energy sector (Su et al. 2021, Zhang et al. 2022, Martin-Valmayor et al. 2023), especially on electricity spot prices. For instance, the stability of oil and gas supplies used for electricity generation has been significantly affected since the outbreak of the Russian-Ukrainian war. The interruption of these commodities may lead to electricity shortages, which results in price volatilities and extreme prices (Saâdaoui and Jabeur 2023). As a result, the need for managing electricity price risks has become even more urgent in increasingly complex electricity markets.

Various methods have been employed for the electricity market to manage price risks, mainly including hedging and forecasting (Deng and Oren 2006, Conejo et al. 2010, Çanakoğlu and Adıyeke 2020, Janczura and Wójcik 2022). However, in many regions, particularly in emerging countries, financial derivatives such as futures and options that are commonly used for hedging price risks are either underdeveloped or unavailable (Avci Surucu et al. 2018). As a result, electricity price forecasting (EPF) has become the primary tool for market participants to manage price risks. Given its importance, substantial research has focused on accurate EPF, with two broad categories of methods: statistical methods and machine learning methods (Nowotarski and Weron 2018, Loi and Le Ng 2018, Chang et al. 2019, Lago et al. 2018). However, accurately forecasting extreme price occurrences remains a significant challenge. Extreme price events are rare, often occur suddenly with large magnitudes, and are typically treated as outliers during data processing. This rarity and unpredictability make them particularly difficult to model effectively (He et al. 2015, Tafakori et al. 2018).

In practice, both market participants and operators need accurate forecasting of extreme price occurrences to address operational challenges arising from price risks in the market (Christensen et al. 2012, Clements et al. 2013). This is because when spot prices exceed or fall below a certain threshold, their trading and operational decisions should be adjusted accordingly. For instance, generators on the sell-side need to avoid underpricing or incurring losses during periods of extremely low prices, while retailers or consumers on the buy-side need to mitigate the risk of overpaying during periods of extremely high prices (Ullah et al. 2018). Moreover, market operators depend on accurate extreme price forecasts to reduce the costs associated with daily operations, such as managing supply-demand imbalances and monitoring potential instances of market power abuse by participants (Westgaard et al. 2021).

Therefore, accurately forecasting extreme price occurrences is essential for managers handling risk management and operations, as well as for regulators overseeing the market in such circumstances.

1.1 | Definition and occurrences of extreme price

The occurrence of extreme prices can be regarded as binary events, and forecasting the occurrence of extreme prices can be formulated as a binary classification problem (Hagfors et al. 2016). Extreme prices can usually be filtered out from normal prices through a threshold, and exceeding the threshold indicates the occurrence of extreme prices. Since extreme prices are not readily observable, the first step is to determine the threshold for differentiating normal and extreme prices. In general, there are two ways in the literature to pre-determine price thresholds: fixed price threshold and variable price threshold. Fixed price threshold means to determine a fixed price value such as 100 Eur/MWh, 120 Eur/MWh or 150 Eur/MWh as threshold (Herrera and González 2014, He and Chen 2016, Clements et al. 2015, Manner et al. 2016, Galarneau-Vincent et al. 2023), whereas variable price threshold determines a certain percentage of the highest (e.g., 99%, 95%, 90%) and the lowest (e.g., 1%, 5%, 10%) prices (Trueck et al. 2007, Sandhu et al. 2016, Liu et al. 2022a). Variable thresholds provide more flexibility than fixed thresholds as they can change with the dynamics of market prices.

However, while variable thresholds introduce adaptability, they rely on a static percentile computed over the entire dataset, which limits their responsiveness to evolving market dynamics. A threshold derived from global statistics fails to capture temporal variations, particularly those induced by significant external shocks. For instance, geopolitical conflicts, extreme weather events, and policy shifts can drive substantial price fluctuations in certain periods while leaving other timeframes

relatively stable. This imbalance skews extreme price classification toward those affected years, potentially overlooking relatively high prices during normal periods.

Thus, given the increasing complexity of renewable electricity markets and the heightened influence of geopolitical events in Europe, existing thresholding methods struggle to define extreme prices effectively. This necessitates a more adaptive approach to identify extreme price occurrences accurately.

1.2 | Methods for forecasting extreme price occurrences

In the literature, many approaches have been studied to forecast the occurrence of extreme prices. Traditionally, statistical approaches, such as autoregressive models and logistic regression models, have been widely employed to model and forecast extreme price occurrences (Eichler et al. 2014, Manner et al. 2016, Maryniak and Weron 2019, Liu et al. 2022a, Adline and Ikeda 2023). However, these econometric approaches often face challenges in capturing the non-linear relationships and complex interactions between features inherent in electricity markets.

Over the past decades, machine learning methods have demonstrated superior performance in handling non-linear relationships and incorporating a large number of predictors (Galarneau-Vincent et al. 2023), making them more suitable for complex electricity market price data. Thus, scholars have increasingly applied machine learning models to forecast extreme price occurrences, achieving notable improvements in forecasting performance. Various machine learning models have been studied in the literature, including random forest (RF) (Datta and Datta 2016, He and Chen 2016, Galarneau-Vincent et al. 2023), deep neural networks (DNNs) (Yamada and Mori 2021, Liu et al. 2022b) and gradient boosting decision tree (GBDT) (Stathakis et al. 2021, Galarneau-Vincent et al. 2023, Zamudio López et al. 2024).

Among these machine learning methods, boosting ensemble approaches, particularly Extreme Gradient Boosting (XGBoost) (Chen and Guestrin 2016), have emerged as state-of-the-art techniques in various predictive tasks. Due to its fast training speed and high predictive accuracy, XGBoost has become a widely preferred model across diverse application domains (Ma et al. 2021, Rawson et al. 2022, Hunt et al. 2022, Zhang et al. 2023, Asselman et al. 2023, Niazkar et al. 2024, Yao et al. 2024). While XGBoost has been applied to normal electricity price forecasting (Galarneau-Vincent et al. 2023), research specifically focusing on forecasting extreme price occurrences remains limited.

1.3 | Challenges on classification of imbalanced data and its interpretability

In practice, occurrences of extremely high or low prices are relatively rare, making them significantly less frequent than normal price events and resulting in a highly imbalanced class dataset. However, standard machine learning algorithms are not inherently optimized for imbalanced datasets in classification problems. In many cases, these models have struggled to deliver satisfactory results when classifying imbalanced data (Ruisen et al. 2018, Averro et al. 2023). Therefore, addressing the class imbalance is critical for achieving effective model performance.

In recent years, several approaches have been proposed to enhance the performance in handling class imbalance. These methods can be broadly categorized into two groups: data-level methods and algorithm-level methods (Tanha et al. 2020). Data-level approaches aim to rebalance the data distribution through resampling techniques, such as random over-sampling, random under-sampling, and the Synthetic Minority Over-sampling Technique (SMOTE) (Del Rio et al. 2015, Dhankhad et al. 2018, Varmedja et al. 2019). However, these methods alter the original data distribution, potentially degrading data quality and leading to overfitting or underfitting issues (Liu et al. 2022c).

Algorithm-level methods primarily adjust the learning process to mitigate bias toward majority classes (Tanha et al. 2020, Wang et al. 2020). A common approach is cost-sensitive learning, where the model adjusts the misclassification cost for different classes, assigning a higher penalty to errors in the minority class. This encourages the model to focus more on correctly classifying minority class instances while still considering majority class performance (Ling and Sheng 2008). The goal is to minimize the overall misclassification cost across the training dataset. However, determining appropriate cost values is challenging, as it requires balancing multiple factors with trade-offs (Tanha et al. 2020).

In contrast to cost-sensitive learning, some machine learning frameworks have integrated class-weighting mechanisms into their algorithms, making them more practical and accessible. Models such as Logistic Regression with class weighting, Random Forest with class weighting, and Weighted-XGBoost (Averro et al. 2023) offer built-in solutions that can effectively handle class imbalance with minimal parameter adjustments.

Given the stringent timeliness and accuracy requirements for forecasting extreme electricity price occurrences, weighted-XGBoost emerges as a particularly suitable choice (Averro et al. 2023). Nevertheless, limited research has explicitly addressed the class imbalance issue in forecasting extreme electricity price occurrences, highlighting an overlooked gap in the literature.

In addition, as machine learning models become more accurate in forecasting the occurrence of extreme prices, they also become less interpretable and are often regarded as black-box models (Machlev et al. 2022). This poses a significant disadvantage for market stakeholders, as they are not only interested in accurately forecasting extreme price occurrences but also in understanding the driving factors behind them. Such insights are crucial for supporting effective price risk management and informed decision-making.

Explainable Artificial Intelligence (xAI) has gained considerable attention due to the increasing demand for interpretability of remarkable black-box models (Molnar 2020, Burkart and Huber 2021). It is a branch of machine learning research that focuses on designing human-understandable models and providing post-modeling explanations for black-box models (Tjoa and Guan 2020). An important subfield of xAI explores the predictability of desired labels or values based on the input features (Lundberg et al. 2020). In this context, SHAP values (SHapley Additive exPlanations) have emerged as a widely used approach to quantify the contribution of each input feature to the output of machine learning models (Lundberg and Lee 2017).

SHAP values have been successfully applied in electricity system research to identify drivers and assess risks related to power grid frequency stability (Kruse et al. 2021ab), explain load forecasting (Lee et al. 2020), gain insights into PV power generation forecasting (Chang et al. 2020, Mitrentsis and Lens 2022), and assess the forecastability of electricity prices (Tschora et al. 2022, Trebbien et al. 2023, Cramer et al. 2023). SHAP values have been extensively utilized in the electricity system, however, the specific challenge of explaining the occurrences of extreme electricity price forecasts has yet to be addressed.

1.4 | Research aims and contributions

To fill the gaps, this research aims to develop an interpretable machine learning framework to forecast imbalanced extreme electricity price occurrences and analyze the contribution of different input features. To achieve this, this study introduces a dynamic weighted threshold method to identify extreme electricity prices, and weighted-XGBoost serves as the primary forecasting model. The framework benchmarks its performance against Logistic Regression and Random Forest models with class weighting. It incorporates historical electricity prices, market characteristics (such as load and generation), fuel prices, and a geopolitical indicator as forecasting features. Finally, it applies SHapley Additive exPlanations (SHAP) to assess feature contributions and identify key drivers of extreme price occurrences.

The key contributions of this paper can be summarised as follows.

- Introduces a dynamic weighted threshold method to identify extreme electricity prices, dynamically adapting to market conditions. Unlike static thresholds, this method ensures robust identification of extreme prices in both crises and stable periods.
- Incorporates the Geopolitical Risk Index as a novel feature in extreme electricity price forecasting, examining its potential impact in the context of geopolitical dynamics.
- Employs a weighted-XGBoost model to improve the forecasting accuracy of rare extreme price events, overcoming the limitations of standard machine learning models in handling imbalanced data.
- Applies SHAP-based explainability to decompose model results and assess feature importance. This provides valuable insights into the mechanisms driving extreme price occurrences, supporting more informed risk management strategies.

The remainder of the paper is organized as follows. Section 2 describes the data and methodological development, including the proposed dynamic weighted threshold and weighted-XGBoost model. Section 3 presents results and discussion, including extreme electricity price forecasts, comparative performance analysis against baseline models, and feature contribution analysis based on SHAP values. Finally, Section 4 provides a summary of the paper along with conclusions.

2 | DATA AND METHODOLOGY

This section consists of three phases, as illustrated in Figure 1. The first phase covers data collection and pre-processing, detailed in Sections 2.1, 2.2, and 2.3. The second phase focuses on machine learning modelling, discussed in Section 2.4. The last phase covers methods of forecast evaluation and interpreting analysis, presented in Sections 2.5 and 2.6.

2.1 | Data

This research focuses primarily on the European electricity spot market. Five years of data from January 1, 2019, to December 31, 2023, are chosen as this period primarily encompasses the European energy crisis. In addition, since Germany's day-ahead electricity market is one of the most actively traded spot markets with a high penetration of renewable energy resources in Europe (Lehna et al. 2022), German data is selected for research. For power market data, day-ahead electricity prices, residual load forecasts, total load forecasts, and

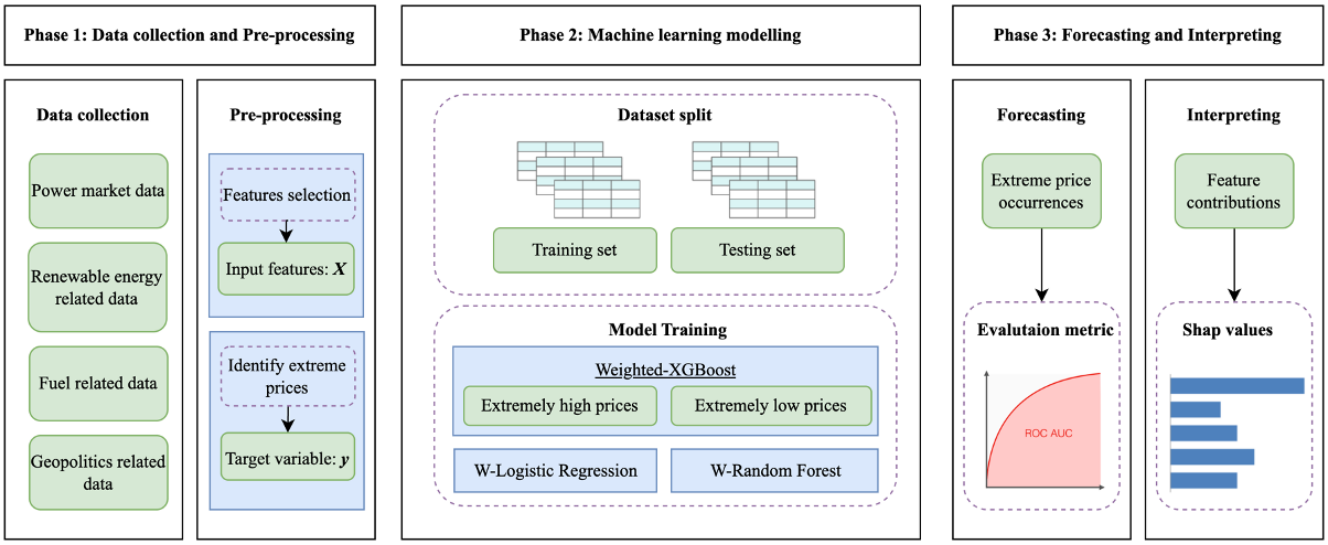


FIGURE 1 Proposed methodology for forecasting and interpreting extreme electricity prices.

total generation forecasts in the German market are obtained from the publicly available German Federal Network Agencys SMARD platform (**?**), as well as renewable energy-related input variables: wind and photovoltaics generation forecasts.

Although renewable energy resources already account for more than half of Germany's electricity production, coal- and natural gas-based power generation continues to be a crucial complement to intermittent generation (Madadkhani and Ikonnikova 2024). Therefore, EGIX (European Gas Index) and API2 (ARA coal price) indexes are considered in this research. To capture the influence of carbon pricing, EUA (European Union Allowance) carbon prices are further included in the set of variables. These daily data are obtained from Refinitiv (2024). Since the market closed on weekends, missing observations are calculated by interpolating the values from the forward-filled approach (i.e., Friday's data) (Mankoff et al. 2020).

Furthermore, this study takes into account the geopolitical impact conditions. Even though the geopolitical impact may already be transmitted to electricity prices through commodity prices, we still include a geopolitical risk index in the model to better understand its impact on the occurrence of extreme electricity prices. This is due to their impact on electricity prices may also be direct and independent. Prior research has suggested that geopolitical risks can significantly affect European electricity prices (Saâdaoui and Jabeur 2023, Abdullah et al. 2023). Thus, the daily geopolitical risk index proposed by Caldara and Iacoviello (2022) is considered in this research.

The lower-frequency daily data is converted to hourly data using a forward-fill approach to align it with hourly observations (Mankoff et al. 2020). This method ensures that all 24 hourly values within the same day remain identical after the conversion. As a result, we obtained 43824 hourly values for the DA electricity price and each variable, corresponding to 1826 days of data (5 years), with 24 observations per day. Table 1 reports the descriptions of the model variables and data sources, and Figure 2 shows plots of the time-dependent data.

2.2 | Input features

To ensure the feasibility of our forecasting approach, all input features must be available at the time of the forecast. Specifically, we forecast extremely high and low electricity prices for a given hour on day d, using only information accessible by the previous day ($d-1$). This ensures that the model remains applicable to real-world forecasting scenarios.

Electricity prices exhibit strong daily and weekly patterns due to fluctuations in supply and demand dynamics (Liu et al. 2022a). Accordingly, historical prices from d–1, which capture short-term daily cyclicality, and $d-7$, which reflect recurring weekly trends, are incorporated as key forecasting features. Additionally, market forecasts such as electricity load and generation forecasts, published on $d-1$, provide essential forward-looking information on expected market conditions. Furthermore, fuel prices and geopolitical indicators are included using the most recent available data ($d-1$), as they offer valuable insights into broader economic and political factors affecting electricity markets. The selected features are categorized as follows:

- **Historical prices:** Day-ahead electricity prices and one week earlier ($P_{\tau,d-7}$).

TABLE 1 Descriptions of the model variables and data sources.

Data	Units	Frequency	Description	Data sources
Power market data				
Day-ahead prices	EUR/MWh	Hourly	Electricity prices in the day-ahead market.	ENTSOE (2024)
Load forecast	MWh	Hourly	Forecasted total electricity demand in the power system.	ENTSOE (2024)
Generation forecast	MWh	Hourly	Forecasted total electricity generation in the power system.	ENTSOE (2024)
Residual load forecast	MWh	Hourly	Forecasted remaining demand that renewable energy cannot cover.	ENTSOE (2024)
Renewable energy-related data				
Wind generation forecast	MWh	Hourly	Forecasted electricity generation from wind power sources.	ENTSOE (2024)
PV generation forecast	MWh	Hourly	Forecasted electricity generation from photovoltaic solar panels.	ENTSOE (2024)
Fuel-related data				
Gas prices	EUR/MWh	Daily	European Gas Index (EGIX), representing natural gas prices.	Refinitiv (2024)
Coal prices	EUR/MWh	Daily	API2 (ARA coal price) index, representing coal prices.	Refinitiv (2024)
Carbon prices	EUR/tCO2	Daily	EUA carbon prices, representing the cost of carbon emissions.	Refinitiv (2024)
Geopolitics-related data				
Geopolitical Risk Index	Point	Daily	A measure of adverse geopolitical events and associated risks.	Caldara and Iacoviello (2022)

- **Load forecasts:** Day-ahead forecasts for total load ($X^{\text{Load}}_{\tau,d}$) and residual load ($X^{\text{Res}}_{\tau,d}$) at hour τ on day d.
- **Generation forecasts:** Day-ahead forecasts for total generation ($X^{\text{Gen}}_{\tau,d}$), wind generation ($X^{\text{Wind}}_{\tau,d}$), and photovoltaic (PV) generation ($X^{\text{PV}}_{\tau,d}$) at hour τ on day d.
- **Fuel-related prices:** Closing prices of key fuel at hour τ on the previous day, including Carbon ($X^{\text{Carbon}}_{\tau,d-1}$), Gas ($X^{\text{Gas}}_{\tau,d-1}$), and Coal ($X^{\text{Coal}}_{\tau,d-1}$).
- **Geopolitical indicators:** Geopolitical Risk Index at hour τ on the previous day ($X^{\text{GPR}}_{\tau,d-1}$).

Here, τ represents the hour of the day, taking values from 0 to 23, and d represents the day, ranging from 1 (January 1, 2019) to 1826 (December 31, 2023). $d-1$ refers to the previous day and $d-7$ corresponds to the one week earlier.

2.3 | Indentifying extreme prices

With the input features established, the next step is to define the target variable for forecasting extreme electricity prices. This study formulates the task as a supervised binary classification problem. Traditional approaches often rely on a static percentile computed over the entire dataset, which fails to adapt to periods of high volatility. To enhance extreme price identification, we propose an adaptive, dynamic weighted threshold (DWT) method, integrating static global quantiles (capturing overall market trends) with rolling local quantiles (reflecting short-term fluctuations). The weights are dynamically adjusted based on market volatility, allowing the thresholds to adapt effectively to price variations.

This approach, inspired by applications of rolling quantile in risk management, volatility forecasting, and trading strategy optimization (Marshall et al. 2017, Packham et al. 2017, Jiang et al. 2019), ensures more responsive classification compared to static quantile methods.

To formalize this method, we define a rolling window $P^{(r)}_{\tau,d}$ over the price time series P. The rolling window at hour τ on day d is defined as:

$$P^{(r)}_{\tau,d} = \{p_{\tau',d'} \mid d' \in [d-r+1, d], \tau' \in [0, 23]\}, \tag{1}$$

where r denotes the window length in days. The set $P^{(r)}_{\tau,d}$ therefore contains all historical hourly prices over the past r days, including every hour τ' of each day d'. This construction ensures that the rolling window captures a complete sequence of past hourly prices across r consecutive days, covering all 24 hours per day.

Extreme price thresholds are determined using both global and local quantiles. Global quantiles capture long-term trends by computing quantiles over the entire time series P:

$$\begin{aligned} \theta^G(\alpha) &= \inf\{p : F(p) \geq \alpha\}, \\ \theta^G(1-\alpha) &= \sup\{p : F(p) \leq 1-\alpha\}, \end{aligned} \tag{2}$$

where $F(p)$ is the cumulative distribution function (CDF) of P, and α is the quantile level, restricted to $\alpha \in (0, 1)$.

Local quantiles, in contrast, adapt to short-term fluctuations by computing quantiles within the rolling window $P^{(r)}_{\tau,d}$:

$$\begin{aligned} \theta^L_{\tau,d}(\alpha) &= \inf\{p : F^{(r)}_{\tau,d}(p) \geq \alpha\}, \\ \theta^L_{\tau,d}(1-\alpha) &= \sup\{p : F^{(r)}_{\tau,d}(p) \leq 1-\alpha\}, \end{aligned} \tag{3}$$

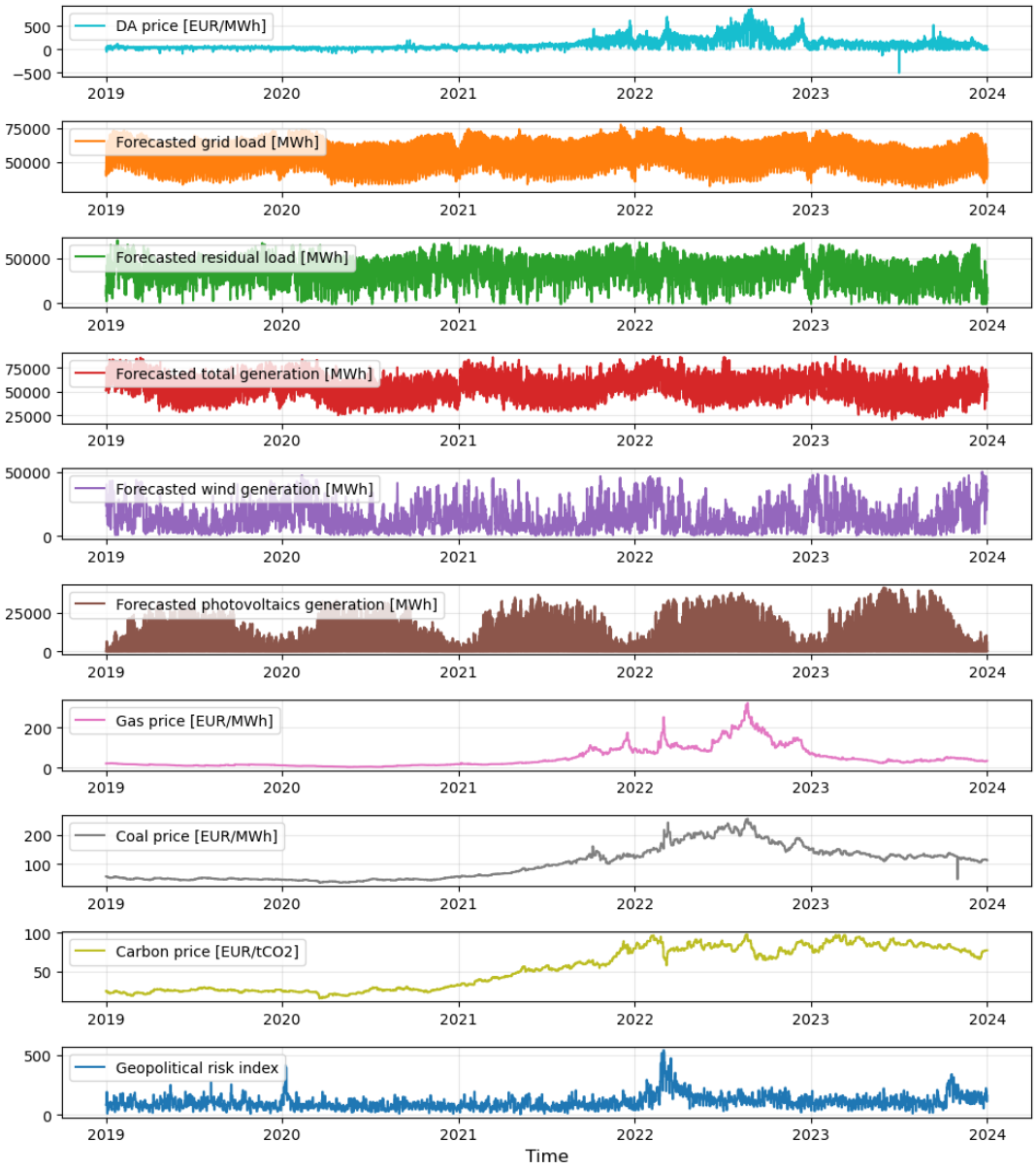


FIGURE 2 Time series plots of the considered data.

where $F^{(r)}_{\tau,d}(p)$ is the CDF of the price set in the rolling window $P^{(r)}_{\tau,d}$.

Volatility is a key indicator of market conditions and structural market changes (Lin and Wesseh Jr 2013, Hernandez et al. 2022). Therefore, we employ it as an adaptive weighting factor to dynamically adjust the balance between global and local quantiles. The local volatility $\sigma_{\tau,d}$ is computed as the standard deviation of prices within $P^{(r)}_{\tau,d}$:

$$\sigma_{\tau,d} = \sqrt{\frac{1}{r \cdot 24} \sum_{p \in P^{(r)}_{\tau,d}} (p - \mu_{\tau,d})^2}, \quad \mu_{\tau,d} = \frac{1}{r \cdot 24} \sum_{p \in P^{(r)}_{\tau,d}} p. \quad (4)$$

To ensure consistency, $\sigma_{\tau,d}$ is normalized between 0 and 1:

$$\sigma^{\text{norm}}_{\tau,d} = \frac{\sigma_{\tau,d} - \min(\sigma)}{\max(\sigma) - \min(\sigma)}. \quad (5)$$

Finally, the dynamic weighted thresholds for extreme price classification are computed as:

$$\begin{aligned} D_{\tau,d}(\alpha) &= \sigma^{\text{norm}}_{\tau,d}\theta^G(\alpha) + (1 - \sigma^{\text{norm}}_{\tau,d})\theta^L_{\tau,d}(\alpha), \\ D_{\tau,d}(1-\alpha) &= \sigma^{\text{norm}}_{\tau,d}\theta^G(1-\alpha) + (1 - \sigma^{\text{norm}}_{\tau,d})\theta^L_{\tau,d}(1-\alpha). \end{aligned} \quad (6)$$

This adaptive weighting scheme follows an intuitive principle: global quantiles are weighted more for threshold stability when local volatility is high, since local price information

is fluctuating and unstable. Conversely, local quantiles are weighted more for better threshold responsiveness when local volatility is low, since local price information is more reliable.

This mechanism ensures that threshold values adjust dynamically without excessive fluctuations, making them more effective for real-world decision-making in risk management. The approach effectively balances stability and adaptability, ensuring that extreme price thresholds remain useful in both calm and volatile market conditions.

Figure 3 illustrates the effectiveness of the DWT method, using a 30-day rolling window as an example, compared to static quantile-based thresholds. The static thresholds (dashed lines) remain fixed at the 90th and 10th quantiles over the entire dataset, failing to adapt to evolving market conditions.

In contrast, the dynamic weighted thresholds (solid lines) adjust in response to market fluctuations. Notably, during periods of extreme market shocks (e.g., post-COVID energy crisis and the Russia-Ukraine war), the dynamic weighted threshold adapts to the changing market while maintaining stability. This adaptive behaviour is essential for extreme price forecasting, ensuring that classification thresholds remain relevant even as market conditions evolve.

Using the dynamic weighted thresholds, we define extreme price occurrences as follows:

$$y_{\tau,d}^{\text{high}} = \begin{cases} 1, & p_{\tau,d} > D_{\tau,d}(\alpha) \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

where $y_{\tau,d}^{\text{high}} = 1$ indicates that $P_{\tau,d}$ is classified as an extremely high price at hour τ on day d, while $y_{\tau,d}^{\text{high}} = 0$ denotes a non-extreme price.

Similarly, extremely low prices are labelled as:

$$y_{\tau,d}^{\text{low}} = \begin{cases} 1, & p_{\tau,d} < D_{\tau,d}(1-\alpha) \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

where $y_{\tau,d}^{\text{low}} = 1$ indicates that $p_{\tau,d}$ is classified as an extremely low price at hour τ on day d, and $y_{\tau,d}^{\text{low}} = 0$ denotes a non-extreme price.

In this paper, extremely high and low prices are classified separately to facilitate targeted forecasting and to examine the distinct mechanisms driving these price extremes. Furthermore, typical values for α are set to 0.9, enabling the identification of the top and bottom prices based on DWT, respectively.

Since the DWT approach integrates both global and local information, it is crucial to use a relatively short window r to effectively capture immediate fluctuations. Given this, we set the window size r to 30 days, corresponding to $30 \times 24 = 720$ hourly observations. In practice, decision-makers may consider different time window sizes to capture recent information. Therefore, we also evaluate the method using varying window sizes ± 15 days, resulting in three window lengths: $15 \times 24 = 360$, $30 \times 24 = 720$, and $45 \times 24 = 1080$ hourly observations. With these thresholds established, extreme price occurrences, serving as the target variables for classification, can be determined. Table 2 summarizes the counts of extreme price occurrences identified using this method.

TABLE 2 Overview of extreme price occurrence counts.

Type of experiments	Type of prices	Time window		
		15-day	30-day	45-day
Extremely high	Extreme prices	4213	4376	4373
	Normal prices	39611	39448	39451
Extremely low	Extreme prices	4047	4053	3995
	Normal prices	39777	39771	39829

2.4 | Weighted-XGBoost classifier

Weighted-XGBoost is an enhanced version of XGBoost designed to address class imbalance issues (Averro et al. 2023). It is integrated into the XGBoost package and can be activated by setting the parameter *scale_pos_weight* during model training (Chen and Guestrin 2016). In this study, this method is tailored to forecast extreme electricity price occurrences, which occur much less frequently than normal prices, leading to a highly imbalanced dataset. This section describes how weighted-XGBoost is employed to forecast extreme price occurrences and elaborates on its mathematical foundation.

2.4.1 | Feature Representation and Temporal Indexing

Weighted-XGBoost is implemented as a binary classification model to forecast the occurrence of extremely high and low electricity prices. The dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}$ consists of n hourly observations recorded in strict chronological order, where $\mathbf{x}_i \in \mathbb{R}^M$ is the feature vector , with M denoting the number of input features, and $y_i \in \{0, 1\}$ is the binary target variable. Each sample, indexed by i, corresponds to an observation at hour τ on day d is represented as $(\mathbf{x}_i, y_i)$. The feature vector $\mathbf{x}_i$ includes the following components as determined in Section 2.2:

$$\begin{aligned} \mathbf{x}_i = \{ & p_{\tau,d-1}, p_{\tau,d-7}, x_{\tau,d}^{\text{Load}}, x_{\tau,d}^{\text{Res}}, x_{\tau,d}^{\text{Gen}}, x_{\tau,d}^{\text{Wind}}, x_{\tau,d}^{\text{PV}}, x_{\tau,d-1}^{\text{Carbon}}, \\ & x_{\tau,d-1}^{\text{Gas}}, x_{\tau,d-1}^{\text{Coal}}, x_{\tau,d-1}^{\text{GPR}} \}. \end{aligned} \tag{9}$$

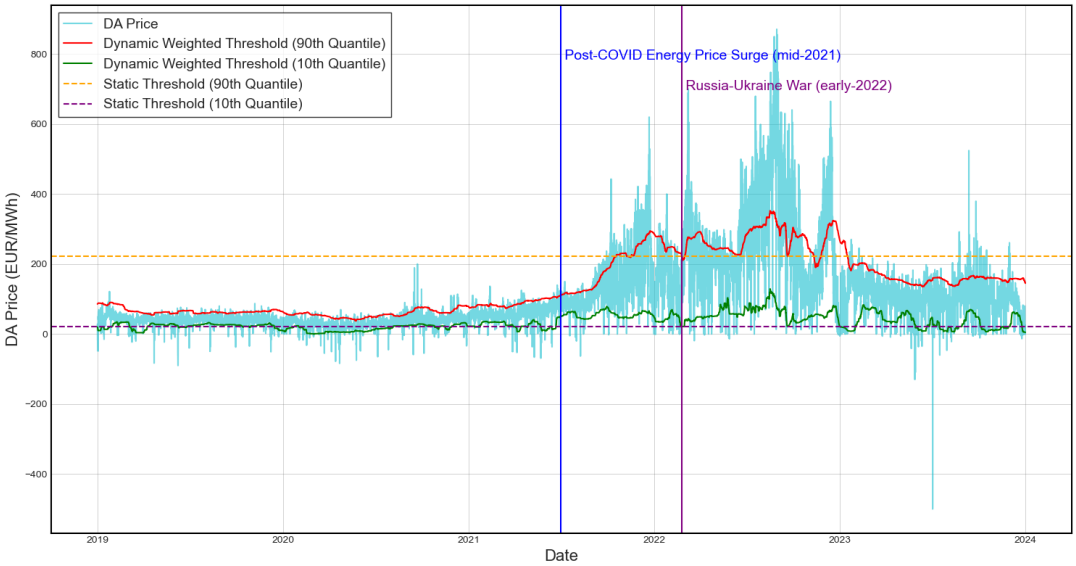


FIGURE 3 Comparison of dynamic weighted thresholds and static thresholds.

The target variable $y_i \in \{0, 1\}$ represents whether an extreme price event occurs at the corresponding time. In this study, two separate weighted-XGBoost models are trained for forecasting extreme price occurrences: one for extremely high ($y_{\tau,d}^{\text{high}}$) and another for extremely low ($y_{\tau,d}^{\text{low}}$), as detailed in Section 2.3.

2.4.2 | Regularized learning objective

In this setup, $(\mathbf{x}_i, y_i)$ represents a data pair, the model's final forecast for the i-th sample, denoted $\hat{y}_i$, can be expressed as shown in Eq. (10):

$$\hat{y}_i = \sum_{k=1}^{K} f_k(\mathbf{x}_i), f_k \in \mathcal{F}, \tag{10}$$

where K denotes the number of trees and $\mathcal{F}$ represents the space of trees utilized in boosting. Each f_k corresponds to an independent tree structure q and leaf weights w. The set of functions f_k can be learned by minimizing the regularized objective function $\mathcal{L}$, as shown in Eq. (11):

$$\mathcal{L} = \sum_{i=1}^{n} l(\hat{y}_i, y_i) + \sum_{k=1}^{K} \Omega(f_k), \tag{11}$$

where $l(\hat{y}_i, y_i)$ represents the binary classification loss function, which quantifies the deviation between the forecasted value $\hat{y}_i$ and the observed value y_i. In classification trees, this loss is typically defined using the logistic loss function, formulated as:

$$l(\hat{y}_i, y_i) = y_i \ln\left(1 + e^{-\hat{y}_i}\right) + (1 - y_i) \ln\left(1 + e^{\hat{y}_i}\right), \tag{12}$$

while $\Omega(f)$ represents the regularization term on decision trees to prevent overfitting:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2, \tag{13}$$

where γT is a structural complexity penalty, with T representing the number of leaf nodes in the tree and γ being a regularization parameter that specifies the minimum loss reduction required to make a further partition on a leaf node. The second term, $\frac{1}{2}\lambda \|w\|^2$, applies l_2 regularization on the leaf weights w, where λ controls the strength of the regularization. This formulation ensures the model does not grow excessively complex while maintaining good generalization performance.

2.4.3 | Gradient tree boosting

Gradient boosting trees adopt an additive modeling approach, where at the t-th iteration, a new decision tree f_t is added to improve the model's forecast $\hat{y}_i^{(t)}$ by minimizing the loss function, as formulated in Eq. (14):

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i)) + \Omega(f_t). \tag{14}$$

To enable efficient optimization of the objective function, the Taylor expansion is applied, resulting in the following expression:

$$\mathcal{L}^{(t)} \backsimeq \sum_{i=1}^{n} \left[l\left(y_i, \hat{y}_i^{(t-1)}\right) + g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(\mathbf{x}_i) \right] + \Omega(f_t), \tag{15}$$

where $g_i = \partial_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right)$ and $h_i = \partial^2_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right)$ represent first- and second- order gradient statistics on the loss function, respectively. The constant term $l\left(y_i, \hat{y}_i^{t-1}\right)$ can be removed to simplify the objective function at step t, shown as follows:

$$\tilde{\mathcal{L}}^{(t)} = \sum_{i=1}^{n} \left[g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(\mathbf{x}_i) \right] + \Omega(f_t). \tag{16}$$

Let $I_j = \{i | q(\mathbf{x}_i) = j\}$ represent the sample set of leaf j. Substituting the expanded form of Ω into the Eq. (16), we obtain the following expression after simplifying:

$$\begin{aligned} \tilde{\mathcal{L}}^{(t)} &= \sum_{i=1}^{n} \left[g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(\mathbf{x}_i) \right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \\ &= \sum_{j=1}^{T} \left[\left(\sum_{i \in I_j} g_i \right) w_j + \frac{1}{2} \left(\sum_{i \in I_j} h_i + \lambda \right) w_j^2 \right] + \gamma T, \end{aligned} \tag{17}$$

where w_j represents the weight of the leaf node j.

The optimal weight w_j^* of leaf j in a fixed structure $q(\mathbf{x})$ can be calculated as following:

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda}, \tag{18}$$

and the corresponding optimal value of the objective function can be found:

$$\tilde{\mathcal{L}}^{(t)}(q) = -\frac{1}{2} \sum_{j=1}^{T} \frac{\left(\sum_{i \in I_j} g_i \right)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T. \tag{19}$$

Eq. (19) can be used to measure the quality of a tree structure q.

However, enumerating all possible tree structures q is computationally infeasible. Instead, a greedy algorithm that iteratively expands the tree by adding branches is employed. At each step, the model evaluates candidate splits and selects the one that maximizes the loss reduction. Given a node containing instance set $I = I_L \cup I_R$, splitting it into left and right nodes, I_L and I_R, respectively, results in the following loss reduction:

$$\mathcal{L}_{\text{split}} = \frac{1}{2} \left[\frac{\left(\sum_{i \in I_L} g_i \right)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{\left(\sum_{i \in I_R} g_i \right)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{\left(\sum_{i \in I} g_i \right)^2}{\sum_{i \in I} h_i + \lambda} \right] - \gamma. \tag{20}$$

If the reduction in loss is positive ($\mathcal{L}_{\text{split}} > 0$), the split is considered beneficial, and the model proceeds with the partition. Otherwise, the split is discarded to prevent unnecessary complexity.

Overall, the proposed tree structure of weighted-XGboost can be established based on the loss function mentioned above.

2.4.4 | Hyperparameter optimization

Furthermore, some parameters cannot be determined through the training process above and must be specified in advance. Machine learning commonly uses hyperparameter optimization techniques to set these parameters in order to achieve the best forecasting performance. Common approaches include grid search, random search, and Bayesian optimization (Yang and Shami 2020, Srinivas and Katarya 2022, Bischl et al. 2023). Grid and random search exhaustively explore the entire parameter space, while Bayesian optimization uses prior information to guide a more efficient search process. By leveraging probabilistic modelling, Bayesian optimization can identify optimal hyperparameters with fewer evaluations (Turner et al. 2021, Stuke et al. 2021, Lindauer et al. 2022). Therefore, this study adopts Bayesian optimization for hyperparameter tuning. Several key hyperparameters are selected for optimizing weighted-XGBoost in this study, including the following:

(1) max_depth: Defines the maximum depth of each tree, controlling the models complexity.
(2) learning rate (η): Determines the step size during optimization, affecting model convergence.
(3) n_estimators: Specifies the number of boosting rounds, balancing model accuracy and computational cost.
(4) gamma (γ): A regularization term that controls tree complexity by setting a minimum loss reduction required for a split.
(5) lambda (λ): Applies an l_2 penalty on leaf weights to prevent overfitting and improve generalization.
(6) scale_pos_weight: Adjusts the penalty for the minority class, enhancing model performance on imbalanced datasets.

2.5 | Model evaluation

For this study's objective of forecasting the occurrence of extreme prices, traditional metrics like Accuracy, Precision, Recall and F1-score provide general insights into overall model performance. However, due to the inherent class imbalance in the data, additional metrics, the geometric mean (G-Mean) and Area Under the Curve (AUC), are particularly critical.

The confusion matrix shown in Table 3 is crucial, and many evaluation metrics are derived from it. It contains the summary of forecasting results of all instances of the dataset used for testing.

Accuracy is the ratio of correctly classified samples to the total number of samples. Precision indicates the proportion of samples predicted as positive by the model that are actually positive. Recall represents the proportion of all actual positive

TABLE 3 Confusion matrix.

Conditions for each sector		**Predicted values**	
		Positive	Negative
Actual values	Positive	True positive (TP)	False negative (FN)
	Negative	False positive (FP)	True negative (TN)

samples that are correctly identified by the model. The F1-score is the harmonic mean of precision and recall. These are shown as:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}, \tag{21}$$

$$Precision = \frac{TP}{TP + FP}, \tag{22}$$

$$Recall = \frac{TP}{TP + FN}, \tag{23}$$

$$F1 - score = \frac{Recall \times Precision \times 2}{Recall + Precision}. \tag{24}$$

In addition, G-Mean balances sensitivity (recall for the minority class) and specificity (recall for the majority class), ensuring that the model performs well across both classes (Zhang et al. 2018). A high G-Mean value indicates that the model is effective in identifying minority class instances without sacrificing accuracy for the majority class.

$$G - Mean = \sqrt{\frac{TP}{TP + FN} \cdot \frac{TN}{TN + FP}}. \tag{25}$$

Furthermore, AUC of the Receiver Operating Characteristic (ROC) is an effective metric for evaluating imbalanced binary classifiers (Zou et al. 2016). ROC curve plots True Positive Rate (TPR) versus False Positive Rate (FPR) at different classification thresholds (Sokolova et al. 2006).

$$\text{TPR} = \frac{TP}{TP + FN}, \tag{26}$$

$$\text{FPR} = \frac{FP}{TN + FP}. \tag{27}$$

AUC values range from 0 to 1, with higher values indicating stronger discriminatory power for both minority and majority classes.

2.6 | SHapley Additive exPlanations (SHAP)

SHAP (Lundberg and Lee 2017) is employed in this study to interpret the predictive model's output. Prior to SHAP, the Shapley value was developed based on game theory (Shapley et al. 1953) as a method for evaluating the contribution of each feature to a models predictions. It quantifies the marginal impact of a feature by analyzing the changes in model output when the feature is included or excluded.

Formally, the Shapley value ϕ_m for feature m is defined as:

$$\phi_m = \sum_{\mathbf{x}_S \subseteq \mathbf{x} \setminus \{m\}} \frac{|\mathbf{x}_S|!(|\mathbf{x}| - |\mathbf{x}_S| - 1)!}{|\mathbf{x}|!} \left[f_{\mathbf{x}_{S\cup\{m\}}}(\mathbf{x}_{S\cup\{m\}}) - f_{\mathbf{x}_S}(\mathbf{x}_S) \right], \tag{28}$$

where $\mathbf{x}$ represents the complete set of input features, and $\mathbf{x}_S$ denotes a subset that excludes feature m. The term $f_{\mathbf{x}_{S\cup\{m\}}}$ refers to the model trained with feature m, while $f_{\mathbf{x}_S}$ corresponds to the model trained without it. Thus, $f_{\mathbf{x}_{S\cup\{m\}}}(\mathbf{x}_{S\cup\{m\}})$ represents the model output when feature m is included, whereas $f_{\mathbf{x}_S}(\mathbf{x}_S)$ represents the output when it is omitted.

However, computing the exact Shapley values is computationally expensive due to the combinatorial complexity involved in estimating all possible feature contributions. To address this, Lundberg and Lee (2017) introduced SHAP as an efficient approximation, leveraging a linear additive explanation model to express feature contributions:

$$g(z') = \phi_0 + \sum_{m=1}^{M} \phi_m z'_m, \tag{29}$$

where M denotes the number of features, as previously defined. z'_m is a binary indicator function that equals 1 if feature m is present and 0 otherwise. The term ϕ_0 represents the expected model output when all features are omitted, serving as a baseline prediction.

By employing SHAP, this study aims to provide a more transparent understanding of how individual features contribute to extreme electricity price occurrences, thereby enhancing the interpretability of the weighted-XGBoost models.

3 | RESULTS AND DISCUSSION

The results on extremely high and extremely low prices with different window sizes r are reported in Section 3.1. A comparative analysis of various machine learning models is presented in Section 3.2. The SHAP-based contribution of various features is discussed in Section 3.3.2.

3.1 | Results on extreme price forecasts

Two types of experiments are conducted to forecast extreme electricity price occurrences, focusing separately on extremely high prices and extremely low prices across various time windows. To evaluate the performance of the machine learning model, the dataset is divided into a training set (75%) and a test set (25%). In both experiments, forecasting is performed

using the proposed weighted-XGBoost with Bayesian optimization. During the hyperparameter optimization phase, the Bayesian optimization algorithm explored different combinations of weighted-XGBoost parameters to maximize the mean AUC score through 10-fold time series cross-validation. After completing the iterative process, the algorithm identifies an optimal set of parameters. The hyperparameter tuning ranges and the optimal parameter results for both experiments are summarized in Table 4. The forecasting results are reported as follows.

The performance of weighted-XGBoost in forecasting extreme electricity prices is summarized in Fig. 4 and Fig. 5, as well as in Table 5 and Table 6, which present the confusion matrices and evaluation metrics across the 15-day, 30-day, and 45-day time windows. The results indicate that the model effectively distinguishes between normal and extreme prices while maintaining a reasonable balance between false positives and false negatives.

For extremely high prices, the model maintains strong classification performance across all time windows. The test Accuracy averages 0.877, while Precision and Recall reach 0.942 and 0.877, respectively, resulting in an average F1-score of 0.898. Furthermore, G-Mean (0.880) and AUC (0.957) confirm the models robustness in handling class imbalance. The ROC curves in Fig. 6 show consistently high test AUC values above 0.95, with scores of 0.954, 0.955, and 0.961 for the 15-day, 30-day, and 45-day windows, respectively. These results demonstrate the strong discriminatory power of weighted-XGBoost in forecasting extremely high price occurrences.

Similarly, for extremely low prices, the model exhibits reliable classification performance. The test Accuracy averages 0.876, with Precision, Recall, and F1-score reaching 0.937, 0.876, and 0.893, respectively. G-Mean (0.908) and AUC (0.968) further validate the models effectiveness in handling class imbalance while maintaining reliable forecasting. As illustrated in the ROC curves in Fig. 7, the AUC values remain consistently high across different time windows, with test AUCs of 0.968, 0.967, and 0.969. Notably, the model exhibits a tiny training-test AUC gap, which remains below 0.02 across all time windows. This suggests that extremely low prices follow clear underlying patterns, making them effective for the model to generalize.

In comparison, a key observation from the results is that forecasting extremely high prices is more challenging than forecasting extremely low prices. This is reflected in the larger discrepancy between training and test AUC scores for extremely high prices, suggesting a higher degree of overfitting. For instance, in the 30-day window, the test AUC for extremely high prices is 0.955, while the training AUC reaches 0.994, resulting in a gap of 0.039. In contrast, for extremely low prices, the gap in the same window is only 0.019. This indicates that extremely low prices exhibit more stable and predictable patterns, whereas extremely high prices may be subject to more volatile external factors, such as supply shortages, demand surges, and geopolitical events.

Despite these challenges, weighted-XGBoost consistently demonstrates strong and reliable performance in forecasting both extremely high and low prices. Across all time windows, test AUC values exceed 0.95, confirming the models robustness. The combination of traditional classification metrics and imbalance-sensitive measures (G-Mean and AUC) ensures a comprehensive evaluation of its capability in handling rare extreme price occurrences. This highlights the models potential as a valuable tool for electricity market participants to manage price risks and improve decision-making in volatile market conditions.

3.2 | Comparative analysis

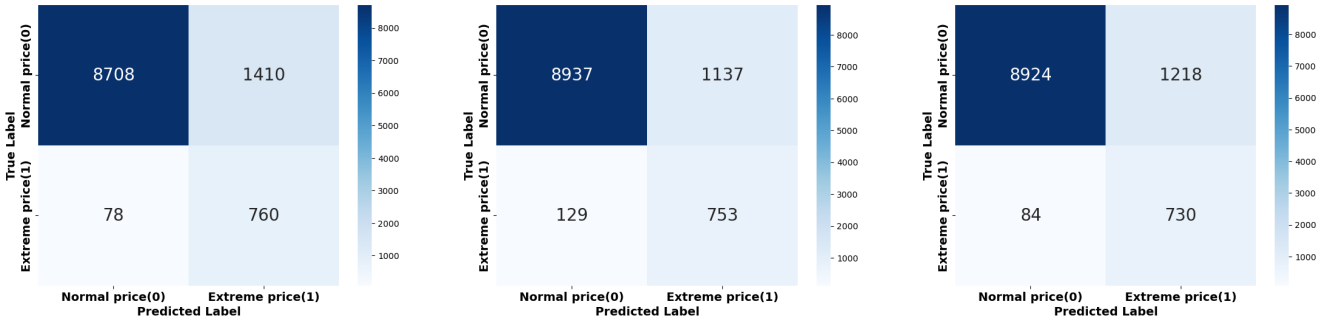
To evaluate the advantages of weighted-XGBoost in forecasting imbalanced extreme electricity prices, the proposed model's performance is compared with two widely used baseline models: Logistic Regression with class weighting (W-LR) and Random Forest with class weighting (W-RF). All models are trained on the same dataset under identical experimental conditions, including the feature set, hyperparameter optimization framework, and evaluation metrics. This ensures that observed performance differences can be attributed to the models' inherent capabilities in handling class imbalance. Two imbalance-sensitive metrics, AUC and G-Mean, are selected to compare the models. AUC provides an overall evaluation of the models discriminatory power, and G-Mean accounts for the balance between sensitivity and specificity, making them well-suited for comparing different models on imbalanced datasets.

The comparative performance of weighted-XGBoost against W-LR and W-RF in forecasting extremely high and low electricity prices is summarized in Table 7 and Table 8, respectively. The results consistently demonstrate the superiority of weighted-XGBoost across all time windows in both AUC and G-Mean metrics for forecasting extremely high and low electricity prices.

For extremely high prices, weighted-XGBoost consistently outperforms the baseline models. It achieves the highest average test AUC of 0.957, compared to 0.941 for W-LR and 0.950 for W-RF. Across different time windows, the test AUC values for weighted-XGBoost remain stable, with scores of 0.954, 0.955, and 0.961 for the 15-day, 30-day, and 45-day windows, respectively. These results highlight the models ability to effectively distinguish between normal and extremely high prices across varying time windows. Similarly, in terms of G-Mean, weighted-XGBoost achieves the highest average test score of

TABLE 4 Hyperparameters tuning for proposed weighted-XGboost in this study.

Hyperparameter	Search range	Type	Optimal value (extremely high prices)			Optimal value (extremely low prices)		
			15-day window	30-day window	45-day window	15-day window	30-day windw	45-day window
learning_rate	[0.005, 0.1]	Real	0.05	0.08	0.05	0.05	0.05	0.08
max_depth	[2, 5]	Integer	4	5	4	4	4	3
n_estimators	[50, 200]	Integer	158	139	158	158	158	91
reg_alpha	[5, 50]	Real	24.09	34.19	24.09	24.09	24.09	19.08
reg_lambda	[5, 50]	Real	34.26	23.99	34.26	34.26	34.26	11.87
gamma	[1, 10]	Real	8.20	6.55	8.20	8.20	8.20	1.03
subsample	[0.5, 1]	Real	0.83	0.82	0.83	0.83	0.83	0.91
min_child_weight	[5, 15]	Integer	14	11	14	14	14	11
max_dalta_step	[5, 10]	Integer	8	7	8	8	8	9
scale_pos_weight	[5, 30]	Real	13.83	23.35	13.83	13.83	13.83	27.35

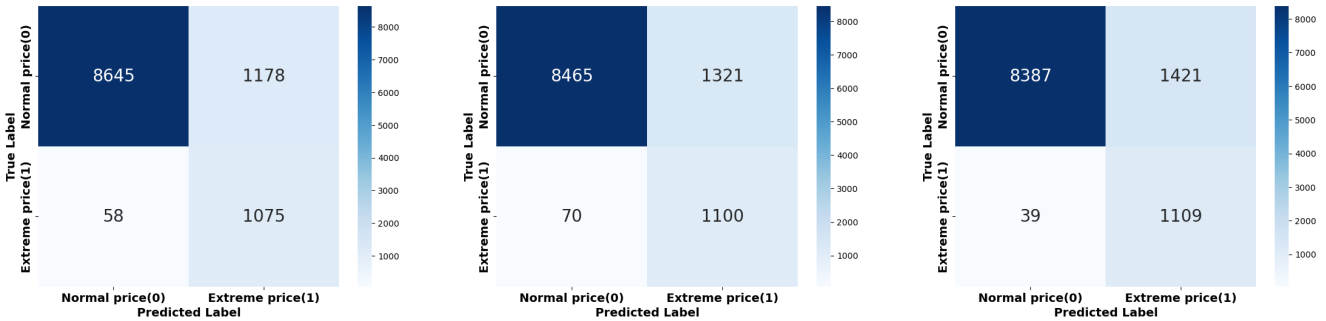


(a) 15-day window **(b)** 30-day window **(c)** 45-day window

FIGURE 4 Confusion matrix for forecasting extremely high prices across different time windows.

0.880, outperforming W-LR (0.793) and W-RF (0.875). The model consistently balances sensitivity and specificity, with test G-Means of 0.883, 0.870, and 0.888 for the 15-day, 30-day, and 45-day windows, respectively.

For extremely low prices, weighted-XGBoost also demonstrates superior performance, achieving the highest scores in both AUC and G-Mean. The model attains an average test AUC of 0.968, surpassing W-LR (0.959) and W-RF (0.962).



(a) 15-day window **(b)** 30-day window **(c)** 45-day window

FIGURE 5 Confusion matrix for forecasting extremely low prices across different time windows.

TABLE 5 Evaluation metric for forecasting extremely high prices using weighted XGBoost across different time windows.

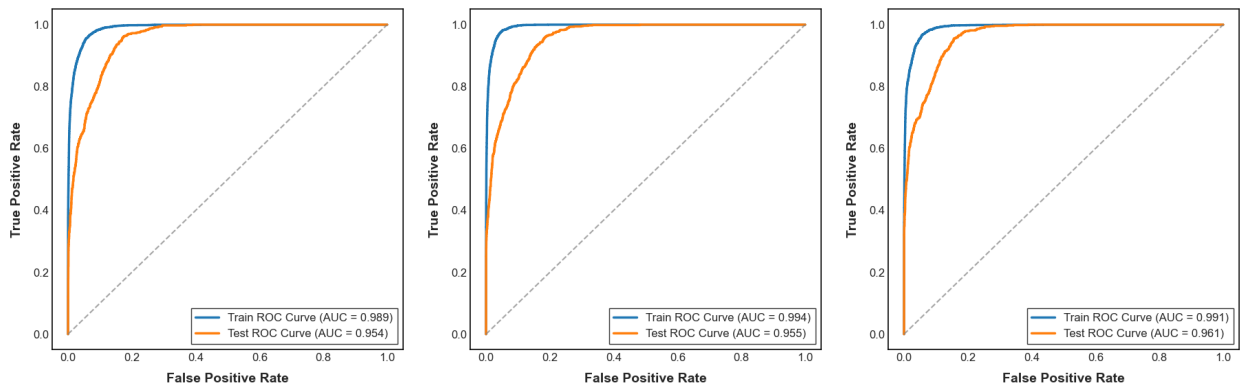
Time window	Data set	Accuracy	Precision	Recall	F1-score	G-Mean	ROC-AUC
15-day	Training	0.919	0.953	0.919	0.928	0.945	0.989
	Test	0.864	0.942	0.864	0.890	0.883	0.954
30-day	Training	0.923	0.955	0.923	0.931	0.953	0.994
	Test	0.885	0.939	0.885	0.902	0.870	0.955
45-day	Training	0.925	0.954	0.925	0.933	0.950	0.991
	Test	0.881	0.945	0.881	0.902	0.888	0.961
Average	Training	0.922	0.954	0.922	0.931	0.949	0.991
	Test	0.877	0.942	0.877	0.898	0.880	0.957

TABLE 6 Evaluation metric for forecasting extremely low prices using weighted XGBoost across different time windows.

Time window	Data set	Accuracy	Precision	Recall	F1-score	G-Mean	ROC-AUC
15-day	Training	0.898	0.950	0.898	0.913	0.934	0.982
	Test	0.887	0.940	0.887	0.902	0.914	0.968
30-day	Training	0.915	0.955	0.915	0.926	0.945	0.986
	Test	0.873	0.934	0.873	0.891	0.902	0.967
45-day	Training	0.891	0.951	0.891	0.908	0.935	0.986
	Test	0.867	0.937	0.867	0.887	0.909	0.969
Average	Training	0.901	0.952	0.901	0.916	0.938	0.985
	Test	0.876	0.937	0.876	0.893	0.908	0.968

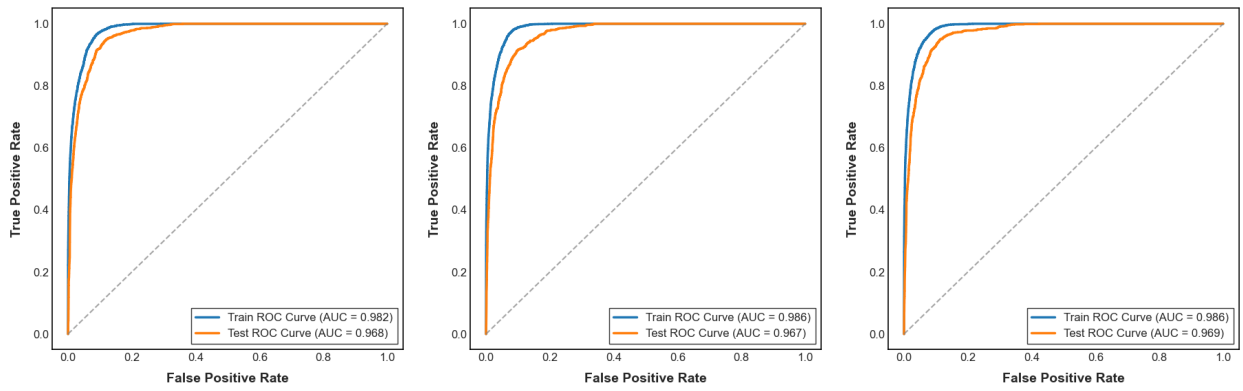
Across the 15-day, 30-day, and 45-day windows, the test AUC values for weighted-XGBoost are 0.968, 0.967, and 0.969, respectively, further confirming its robustness in distinguishing normal and extremely low prices. In terms of G-Mean, weighted-XGBoost again outperforms the baselines, achieving an average test score of 0.908 compared to 0.849 for W-LR and 0.902 for W-RF. For individual time windows, the model attains test G-Means of 0.914, 0.902, and 0.909 for the 15-day, 30-day, and 45-day windows, respectively.

Overall, the results indicate that weighted-XGBoost provides more reliable and accurate forecasts of both extremely high and low prices compared to the baseline models. The consistently higher AUC and G-Mean values suggest that weighted-XGBoost is better suited for handling imbalanced



(a) 15-day window **(b)** 30-day window **(c)** 45-day window

FIGURE 6 ROC curves for forecasting extremely high prices across different time windows.



(a) 15-day window **(b)** 30-day window **(c)** 45-day window

FIGURE 7 ROC curves for forecasting extremely low prices across different time windows.

classification tasks in extreme electricity price forecasting. These findings underscore its potential as a robust and effective tool for market participants seeking to manage price risks and improve forecasting accuracy in volatile electricity markets.

3.3 | Model interpretation

Interpreting the features driving the model's forecasts is essential for validating its effectiveness and ensuring practical applicability, which is crucial for effective price risk management in electricity markets. While performance metrics provide an overall evaluation, they lack insight into the specific contributions of individual features. SHAP (SHapley Additive exPlanations) is employed to address this, offering a transparent view of how features influence the model's output.

SHAP-based analysis can be conducted at both global and local levels to provide a comprehensive understanding of the models decision-making process. The global analysis examines the overall impact of each feature on the model's forecasts by aggregating SHAP values across the entire dataset (Lundberg and Lee 2017). This is typically visualized using SHAP summary plots, which rank features by their average contribution to the forecasted outcome. These plots help identify which factors are most influential in determining extreme electricity prices.

In contrast, the local analysis focuses on explaining individual forecasts by illustrating how specific feature values contribute to a particular forecast (Lundberg et al. 2020). This is achieved through SHAP force plots, which show how different features push a single forecast higher or lower relative to a baseline value, which is the model's expected output before considering any feature influences. These plots offer a detailed breakdown of feature contributions, helping to verify whether extreme price forecasts are driven by meaningful factors.

The following sections present global and local SHAP-based analyses for forecasting extremely high and low electricity prices, highlighting key features and their roles across different time windows.

3.3.1 | Global feature interpretation

The SHAP summary plots in Fig. 8 and Fig. 9 illustrate the global feature importance for forecasting extremely high and low electricity prices across different time windows. These plots provide insights into how individual features influence model forecasts, highlighting the factors that drive extreme price events.

For extremely high prices, multiple interrelated market factors contribute significantly to price surges. Forecasted residual load consistently emerges as the most influential feature across all time windows, reflecting its strong correlation with supply-demand imbalances. High residual load, reflecting unmet demand by renewables and reliance on expensive dispatchable sources, often leads to price spikes under the merit order principle (Trebbien et al. 2023). Fuel-related features, such as coal prices and gas prices, also play a crucial role in shaping high-price occurrences, highlighting the impact of fossil fuel cost fluctuations on electricity markets. Additionally, historical electricity price patterns over daily and weekly periods exhibit significant importance, suggesting that past market trends provide valuable forecasting signals for extremely high prices. The results confirm that extremely high price occurrences are influenced by a combination of supply-demand conditions, fuel price volatility, and historical market behaviours, making their forecasting more challenging and sensitive to external shocks.

For extremely low prices, forecasted residual load exhibits the dominant SHAP values across all time windows, highlighting its central role in driving extreme price decreases. Low

TABLE 7 Comparative analysis for forecasting extremely high prices.

Type of experiment	Model	Data set	Time window			
			15-day	30-day	45-day	Average
AUC	W-LR	Training	0.960	0.965	0.965	0.963
		Test	0.926	0.942	0.954	0.941
	W-RF	Training	0.978	0.978	0.979	0.978
		Test	0.941	0.954	0.956	0.950
	W-XGB	Traning	**0.989**	**0.994**	**0.991**	**0.991**
		Test	**0.954**	**0.955**	**0.961**	**0.957**
G-Mean	W-LR	Training	0.892	0.901	0.900	0.898
		Test	0.772	0.788	0.819	0.793
	W-RF	Training	0.922	0.925	0.927	0.925
		Test	0.861	**0.884**	0.880	0.875
	W-XGB	Training	**0.945**	**0.953**	**0.950**	**0.949**
		Test	**0.883**	0.870	**0.888**	**0.880**

Abbreviations: W-LR = Logistic Regression with class weighting; W-RF = Random Forest with class weighting; W-XGB = Proposed weighted-XGboost.

TABLE 8 Comparative analysis for forecasting extremely low prices.

Type of experiment	Model	Data set	Time window			
			15-day	30-day	45-day	Average
AUC	W-LR	Training	0.940	0.973	0.958	0.957
		Test	0.954	0.966	0.957	0.959
	W-RF	Training	0.967	0.979	0.982	0.976
		Test	0.961	0.963	0.963	0.962
	W-XGB	Traning	**0.982**	**0.986**	**0.986**	**0.985**
		Test	**0.968**	**0.967**	**0.969**	**0.968**
G-Mean	W-LR	Training	0.856	0.913	0.883	0.884
		Test	0.846	0.860	0.841	0.849
	W-RF	Training	0.906	0.925	0.933	0.921
		Test	0.902	0.899	0.905	0.902
	W-XGB	Training	**0.934**	**0.945**	**0.935**	**0.938**
		Test	**0.914**	**0.902**	**0.909**	**0.908**

Abbreviations: W-LR = Logistic Regression with class weighting; W-RF = Random Forest with class weighting; W-XGB = Proposed weighted-XGboost.

or even negative residual load suggests that renewable generation nearly satisfies or exceeds total demand, often reflecting a system-wide oversupply condition. This leads to market-clearing price collapses under the merit-order mechanism. While secondary features, such as carbon price, coal price, and forecasted total generation, also influence forecasts, their SHAP values remain considerably lower compared to the forecasted residual load. Additionally, renewable energy sources, particularly high forecasted wind generation, are observed to exert a downward pressure on electricity prices, reinforcing the role of renewables in shaping price fluctuations. Thus, extremely low prices follow a straightforward and deterministic pattern, primarily governed by supply-demand conditions.

Overall, the comparison between extremely high and low price interpretations reveals notable disparities. Extremely high prices are influenced by a diverse set of factors, including

fuel price volatility, historical price trends, and grid conditions, making them more susceptible to external shocks. In contrast, extremely low prices are largely dictated by forecasted residual load, suggesting a more deterministic relationship between supply-demand conditions and price drops. This distinction also influences the model's generalization ability: forecasts of extremely low prices exhibit greater stability, as indicated by the dominance of a single feature, whereas extremely high price forecasts are more volatile due to the interplay of multiple driving forces.

Additionally, fossil fuel prices exhibit an asymmetric influence on extreme prices. While coal, gas, and carbon prices significantly impact extremely high prices, their effect on extremely low prices is relatively minor. This indicates that high prices are more sensitive to fuel price fluctuations and external market events, such as supply shortages or geopolitical disruptions, whereas low prices primarily result from foreseeable system-wide oversupply conditions.

These insights offer valuable implications for electricity market participants for effective market risk management. The complexity of extremely high prices necessitates dynamic market monitoring and rapid-response strategies to mitigate unexpected price spikes. Meanwhile, the strong forecastability of extremely low prices suggests that system participants can proactively mitigate the impact of electricity oversupply on market stability. By understanding the distinct driving forces behind extreme electricity prices, stakeholders can develop more effective forecasting strategies and market interventions to minimize risks and optimize decision-making.

3.3.2 | Local feature interpretation

While global SHAP analysis provides an overall ranking of feature importance, local SHAP force plots offer case-specific explanations, illustrating how individual feature values contribute to extreme price forecasts. Fig. 10 and Fig. 11 present these force plots for forecasting extremely high and low prices across the 15-day, 30-day, and 45-day time windows. To better understand the models decision-making process, we randomly select instances forecasted as extreme prices and analyze their feature contributions. Each force plot includes a tabular summary displaying four key features that have consistently shown significant influence across different time windows. These features visually depict how different factors push forecasts above or below the base value, ultimately determining whether an extreme price event is forecasted.

The SHAP force plots in Fig. 10 highlight the dynamic interactions between multiple market factors that collectively push forecasts beyond the base value, signaling a higher probability of extremely high price occurrences. In the 15-day window example, high forecasted residual load (4.698×10^4 MWh) plays a dominant role in increasing the models forecasted probability of extreme prices. Coal price and forecasted grid load further reinforce this effect, while lower gas price and day-ahead price exert a negative influence, though to a lesser extent. Similarly, in the 30-day window example, high forecasted residual load (6.112×10^4 MWh), weekly historical price, gas price, and carbon price emerge as key contributors, highlighting the combined effect of past electricity market trends and fossil fuel price fluctuations. In the 45-day window example, high forecasted residual load (5.655×10^4 MWh), forecasted wind generation and electricity price trends gain cole prominence, reinforcing the idea that extremely high price is affected by multiple market factors.

Conversely, the SHAP force plots in Fig. 11 demonstrate a significantly different pattern for forecasting extremely low prices. Across all time windows, low forecasted residual load is consistently the strongest driver of extremely low price occurrences. In the 15-day window example, a drop in residual load (1.188×10^4 MWh) has the greatest positive impact on pushing the models forecast above the base value, indicating a high probability of price dips. While factors such as coal price, carbon price, and gas price exhibit some influence, their impact remains minor. This trend continues in the 30-day and 45-day window examples, where low forecasted residual load remains the primary determining factor (0.9423×10^4 MWh and 1.4830×10^4 MWh), with high wind generation becoming more prominent as a secondary influence.

Overall, the local SHAP force plots reveal a fundamental distinction between extremely high and low price occurrences. Extremely high prices are driven by a broader set of market factors, including fossil fuel prices, historical market trends, and grid load fluctuations, making them more complex and sensitive to external shocks. In contrast, extremely low prices are predominantly dictated by a single factor, forecasted residual load, suggesting that their occurrence follows a more predictable pattern based on supply-demand imbalances.

These insights reinforce the importance of feature-specific risk monitoring in electricity markets. While high prices require tracking multiple economic and operational indicators, low price risks can be primarily managed through real-time monitoring of residual load and renewable energy forecasts.

4 | CONCLUSION

This study explored the application of weighted-XGBoost for forecasting extreme electricity prices in the context of imbalanced data, focusing on both extremely high and low price occurrences. The experimental results demonstrated the robustness and effectiveness of the proposed approach across different time windows, with consistently strong performance

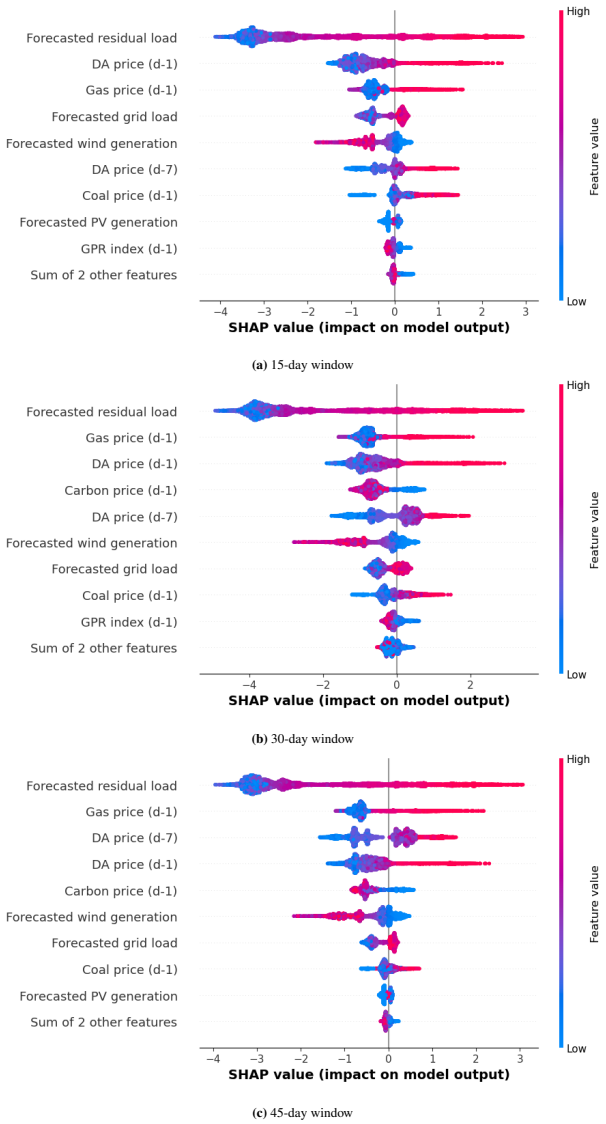


FIGURE 8 SHAP summary plots for extremely high prices across different time windows.

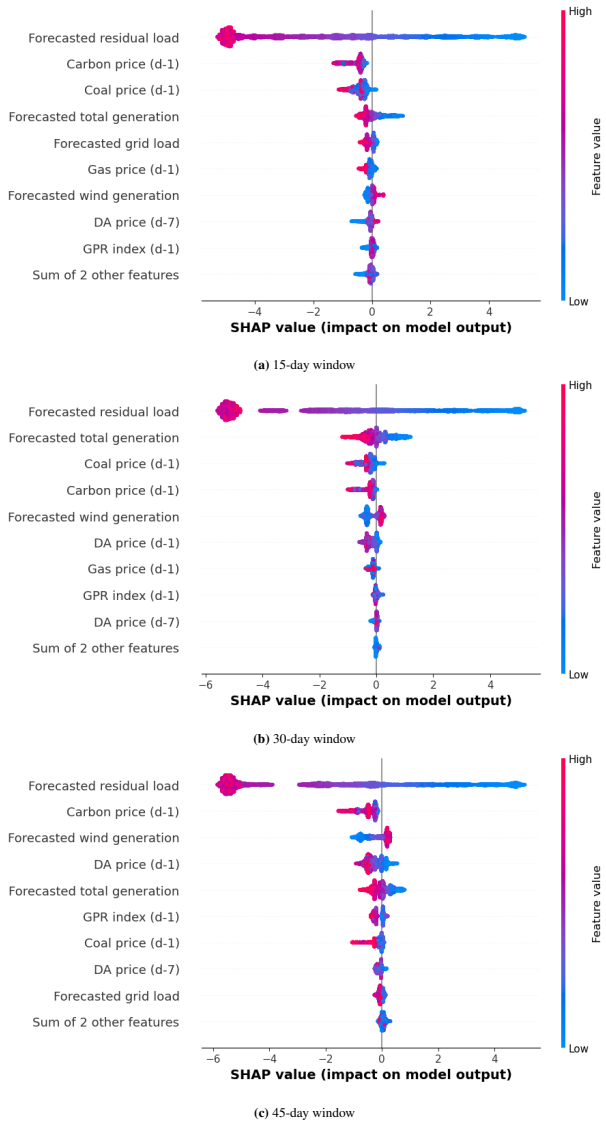


FIGURE 9 SHAP summary plots for extremely low prices across different time windows.

in key metrics such as accuracy, precision, recall, F1-score, G-Mean, and AUC.

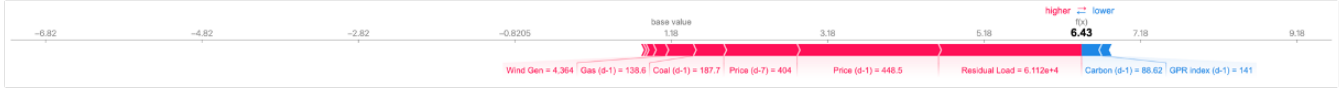
The G-Mean and ROC-AUC metrics confirmed the models ability to effectively distinguish between extreme and normal prices despite the inherent class imbalance. AUC values consistently exceeded 0.95 across training and test datasets, underscoring the models high classification quality. Furthermore, compared to baseline models, weighted-XGBoost achieved superior performance, demonstrating higher forecasting accuracy and better robustness in handling class imbalance. This highlights its effectiveness in capturing the underlying patterns of extreme price occurrences.

The SHAP-based interpretability analysis provided further insights into the key drivers of extreme electricity prices. The results revealed distinct differences between the determinants of extremely high and extremely low prices. Extremely high prices arise from a complex interaction of supply-demand conditions, historical price trends, and external market dynamics, making them more volatile and susceptible to external shocks. Forecasted residual load, grid load, and wind generation play crucial roles, alongside fuel price factors such as gas, carbon, and coal prices, indicating a significant influence of fuel cost fluctuations. In contrast, extremely low prices are primarily driven by forecasted residual load. The lack of strong contributions from price-related variables suggests that low-price events are more structurally determined by demand-side imbalances and renewable energy availability rather than fuel prices and external market shocks.

Sample	Actual class	Predicted class	Base value	Forecasted residual load	DA price (d-1)	DA price (d-7)	Gas price (d-1)
#5310	Normal price (0)	Extreme price (1)	0.4973	46980	116.3	124.1	33.42

(a) 15-day window

Sample	Actual class	Predicted class	Base value	Forecasted residual load	DA price (d-1)	DA price (d-7)	Gas price (d-1)
#1656	Extreme price (1)	Extreme price (1)	1.18	61120	448.5	404	138.6



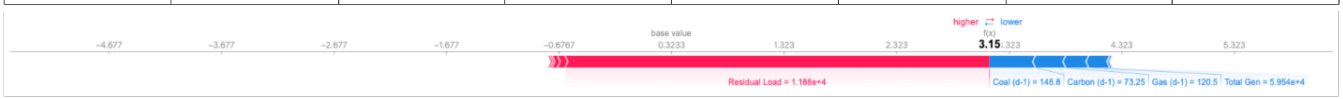
(b) 30-day window

Sample	Actual class	Predicted class	Base value	Forecasted residual load	DA price (d-1)	DA price (d-7)	Gas price (d-1)
#10159	Extreme price (1)	Extreme price (1)	0.5613	56550	117.1	132.3	44.17

(c) 45-day window

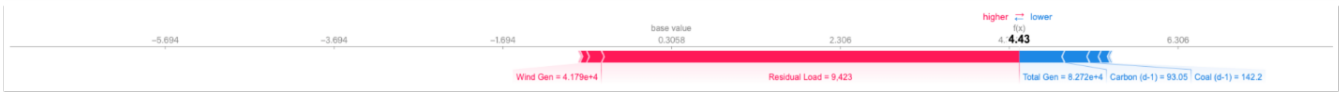
FIGURE 10 SHAP force plots for extremely high prices across different time windows.

Sample	Actual class	Predicted class	Base value	Forecasted residual load	Coal price (d-1)	Carbon price (d-1)	Forecasted total generation
#1120	Extreme price (1)	Extreme price (1)	0.3233	11880	148.8	73.25	59540



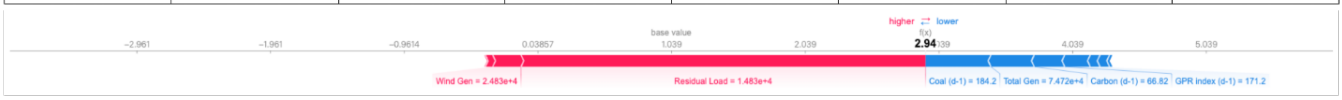
(a) 15-day window

Sample	Actual class	Predicted class	Base value	Forecasted residual load	Coal price (d-1)	Carbon price (d-1)	Forecasted total generation
#3410	Extreme price (1)	Extreme price (1)	0.3058	9423	142.2	93.05	82720



(b) 30-day window

Sample	Actual class	Predicted class	Base value	Forecasted residual load	Coal price (d-1)	Carbon price (d-1)	Forecasted total generation
#120	Extreme price (1)	Extreme price (1)	1.039	14830	184.2	66.82	74720



(c) 45-day window

FIGURE 11 SHAP force plots for extremely low prices across different time windows.

Notably, the geopolitical risk index (GPR) was included as a feature to capture broader economic and political disruptions, yet its SHAP values remained consistently low, indicating minimal influence on extreme electricity price occurrences. This suggests that while geopolitical events may impact long-term energy trends, they do not have a direct, immediate effect on short-term price extremes in the German electricity market.

While the proposed model demonstrates strong forecasting capabilities, there are areas for further improvement. First, this study defines extreme price events using the 90th and 10th quantiles, but future research could explore stricter thresholds (e.g., 95th/5th or 99th/1st percentiles) to assess their impact on forecasting accuracy. Second, while the volatility-based dynamic weighted threshold adapts to market fluctuations, alternative methods such as adaptive quantiles, distribution-based approaches, or extreme value theory (EVT) might refine extreme price identification. Third, while the overall classification performance for forecasting extremely high prices is satisfactory, there is still room for improvement compared to extremely low price forecasts. From the SHAP-based interpretability analysis, it is evident that certain external features play a more significant role in forecasting extremely high prices. This suggests that incorporating additional external features, potentially those more directly relevant to extreme price events, could further enhance the model's ability to accurately forecast extremely high prices. Future work could focus on identifying and integrating such external features to optimize the model's forecasting power for these events.

By addressing these areas, future studies can provide a more comprehensive understanding of extreme price dynamics and improve the robustness of forecasting models in increasingly complex electricity markets. These enhancements would not only refine the methodology but also support market participants and policymakers in better managing risks associated with extreme electricity price occurrences. It is also worth noting that the findings of this study are not limited to the German electricity market. While external conditions and shocks may vary across regions and markets, the methodology proposed in this study is generalizable and can be applied to electricity markets in other regions. This adaptability underscores the broader relevance of the proposed framework for addressing extreme price forecasting challenges in diverse market environments.

REFERENCES

Abdullah, M., Abakah, E.J.A., Ullah, G.W., Tiwari, A.K. & Khan, I. (2023) Tail risk contagion across electricity markets in crisis periods. *Energy Economics*, 127, 107100.

Adline, B. & Ikeda, K. (2023) A hawkes model approach to modeling price spikes in the japanese electricity market. *Energies*, 16(4), 1570.

Asselman, A., Khaldi, M. & Aammou, S. (2023) Enhancing the prediction of student performance based on the machine learning xgboost algorithm. *Interactive Learning Environments*, 31(6), 3360–3379.

Avci Surucu, E., Ketter, W. & van Heck, E. (2018) Managing electricity price modeling risk via ensemble forecasting. *Energy Policy*, 123, 390–403.

Averro, N.T., Murfi, H. & Ardaneswari, G. The imbalance data handling of xgboost in insurance fraud detection. In: *DATA, 2023*, pp. 460–467.

Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S. et al. (2023) Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 13(2), e1484.

Burkart, N. & Huber, M.F. (2021) A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 70, 245–317.

Caldara, D. & Iacoviello, M. (2022) Measuring geopolitical risk. *American Economic Review*, 112(4), 1194–1225.

Çanakoğlu, E. & Adıyeke, E. (2020) Comparison of electricity spot price modelling and risk management applications. *Energies*, 13(18), 4698.

Chang, X., Li, W., Ma, J., Yang, T. & Zomaya, A.Y. Interpretable machine learning in sustainable edge computing: A case study of short-term photovoltaic power output prediction. In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020*, pp. 8981–8985.

Chang, Z., Zhang, Y. & Chen, W. (2019) Electricity price prediction based on hybrid model of adam optimized lstm neural network and wavelet transform. *Energy*, 187, 115804.

Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016*, pp. 785–794.

Christensen, T.M., Hurn, A.S. & Lindsay, K.A. (2012) Forecasting spikes in electricity prices. *International Journal of Forecasting*, 28(2), 400–411.

Clements, A., Fuller, J. & Hurn, S. (2013) Semi-parametric forecasting of spikes in electricity prices. *Economic Record*, 89(287), 508–521.

Clements, A.E., Herrera, R. & Hurn, A. (2015) Modelling interregional links in electricity price spikes. *Energy Economics*, 51, 383–393.

Conejo, A.J., Nogales, F.J., Carrión, M. & Morales, J.M. (2010) Electricity pool prices: long-term uncertainty characterization for futures-market trading and risk management. *Journal of the Operational Research Society*, 61(2), 235–245.

Cramer, E., Witthaut, D., Mitsos, A. & Dahmen, M. (2023) Multivariate probabilistic forecasting of intraday electricity prices using normalizing flows. *Applied Energy*, 346, 121370.

Datta, A.R. & Datta, S. Electricity market price-spike classification in the smart grid. In: *2016 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT). IEEE, 2016*, pp. 1–5.

Del Rio, S., Benítez, J.M. & Herrera, F. Analysis of data preprocessing increasing the oversampling ratio for extremely imbalanced big data classification. In: *2015 IEEE Trustcom/BigDataSE/ISPA*. Vol. 2. *IEEE, 2015*, pp. 180–185.

Deng, S.J. & Oren, S.S. (2006) Electricity derivatives and risk management. *Energy*, 31(6-7), 940–953.

Dhankhad, S., Mohammed, E. & Far, B. Supervised machine learning algorithms for credit card fraudulent transaction detection: a comparative study. In: *2018 IEEE International Conference on Information Reuse and Integration (IRI). IEEE, 2018*, pp. 122–125.

Eichler, M., Grothe, O., Manner, H. & Tuerk, D. (2014) Models for short-term forecasting of spike occurrences in australian electricity markets: a comparative study. *Journal of Energy Markets*, 7(1).

ENTSOE (2024) *Entsoe*. Accessed: 2024-05-23.
URL https://transparency.entsoe.eu/

Galarneau-Vincent, R., Gauthier, G. & Godin, F. (2023) Foreseeing the worst: Forecasting electricity dart spikes. *Energy Economics*, 119, 106521.

Hagfors, L.I., Kamperud, H.H., Paraschiv, F., Prokopczuk, M., Sator, A. & Westgaard, S. (2016) Prediction of extreme price occurrences in the german day-ahead electricity market. *Quantitative Finance*, 16(12), 1929–1948.

He, D. & Chen, W.P. A real-time electricity price forecasting based on the spike clustering analysis. In: *2016 IEEE/PES Transmission and Distribution Conference and Exposition (T&D). IEEE, 2016*, pp. 1–5.

He, K., Yu, L. & Tang, L. (2015) Electricity price forecasting with a bed (bivariate emd denoising) methodology. *Energy*, 91, 601–609.

Hernandez, J.A., Shahzad, S.J.H., Sadorsky, P., Uddin, G.S., Bouri, E. & Kang, S.H. (2022) Regime specific spillovers across us sectors and the role of oil price volatility. *Energy Economics*, 107, 105834.

Herrera, R. & González, N. (2014) The modeling and forecasting of extreme events in electricity spot markets. *International Journal of Forecasting*, 30(3), 477–490.

Hunt, K., Agarwal, P. & Zhuang, J. (2022) Monitoring misinformation on twitter during crisis events: a machine learning approach. *Risk Analysis*, 42(8), 1728–1748.

Janczura, J. & Wójcik, E. (2022) Dynamic short-term risk management strategies for the choice of electricity market based on probabilistic forecasts of profit and risk measures. the german and the polish market case study. *Energy Economics*, 110, 106015.

Jiang, H., Todorova, N., Roca, E. & Su, J.J. (2019) Agricultural commodity futures trading based on cross-country rolling quantile return signals. *Quantitative Finance*, 19(8), 1373–1390.

Ketterer, J.C. (2014) The impact of wind power generation on the electricity price in germany. *Energy Economics*, 44, 270–280.

Kruse, J., Schäfer, B. & Witthaut, D. Exploring deterministic frequency deviations with explainable ai. In: *2021 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm). IEEE, 2021a*, pp. 133–139.

Kruse, J., Schäfer, B. & Witthaut, D. (2021) Revealing drivers and risks for power grid frequency stability with explainable ai. *Patterns*, 2(11).

Lago, J., De Ridder, F. & De Schutter, B. (2018) Forecasting spot electricity prices: Deep learning approaches and empirical comparison of traditional algorithms. *Applied Energy*, 221, 386–405.

Lee, Y., Oh, J. & Kim, G. (2020) Interpretation of load forecasting using explainable artificial intelligence techniques. *The Transactions of the Korean Institute of Electrical Engineers*, 69(3), 480–485.

Lehna, M., Scheller, F. & Herwartz, H. (2022) Forecasting day-ahead electricity prices: A comparison of time series and neural network models taking external regressors into account. *Energy Economics*, 106, 105742.

Lin, B. & Wesseh Jr, P.K. (2013) What causes price volatility and regime shifts in the natural gas market. *Energy*, 55, 553–563.

Lindauer, M., Eggensperger, K., Feurer, M., Biedenkapp, A., Deng, D., Benjamins, C. et al. (2022) Smac3: A versatile bayesian optimization package for hyperparameter optimization. *Journal of Machine Learning Research*, 23(54), 1–9.

Ling, C.X. & Sheng, V.S. (2008) Cost-sensitive learning and the class imbalance problem. *Encyclopedia of Machine Learning*, 2011, 231–235.

Liu, L., Bai, F., Su, C., Ma, C., Yan, R., Li, H. et al. (2022) Forecasting the occurrence of extreme electricity prices using a multivariate logistic regression model. *Energy*, 247, 123417.

Liu, S., Jiang, Y., Lin, Z., Wen, F., Ding, Y. & Yang, L. (2022) Data-driven two-step day-ahead electricity price forecasting considering price spikes. *Journal of Modern Power Systems and Clean Energy*, 11(2), 523–533.

Liu, W., Fan, H., Xia, M. & Pang, C. (2022) Predicting and interpreting financial distress using a weighted boosted tree-based tree. *Engineering Applications of Artificial Intelligence*, 116, 105466.

Loi, T.S.A. & Le Ng, J. (2018) Anticipating electricity prices for future needs–implications for liberalised retail markets. *Applied Energy*, 212, 244–264.

Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B. et al. (2020) From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2(1), 56–67.

Lundberg, S.M. & Lee, S.I. (2017) A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.

Ma, M., Zhao, G., He, B., Li, Q., Dong, H., Wang, S. et al. (2021) Xgboost-based method for flash flood risk assessment. *Journal of Hydrology*, 598, 126382.

Machlev, R., Heistrene, L., Perl, M., Levy, K., Belikov, J., Mannor, S. et al. (2022) Explainable artificial intelligence (xai) techniques for energy and power systems: Review, challenges and opportunities. *Energy and AI*, 9, 100169.

Madadkhani, S. & Ikonnikova, S. (2024) Toward high-resolution projection of electricity prices: A machine learning approach to quantifying the effects of high fuel and co2 prices. *Energy Economics*, 129, 107241.

Maniatis, G.I. & Milonas, N.T. (2022) The impact of wind and solar power generation on the level and volatility of wholesale electricity prices in greece. *Energy Policy*, 170, 113243.

Mankoff, K.D., Solgaard, A., Colgan, W., Ahlstrøm, A.P., Khan, S.A. & Fausto, R.S. (2020) Greenland ice sheet solid ice discharge from 1986 through march 2020. *Earth System Science Data*, 12(2), 1367–1383.

Manner, H., Türk, D. & Eichler, M. (2016) Modeling and forecasting multivariate electricity price spikes. *Energy Economics*, 60, 255–265.

Marshall, B.R., Nguyen, N.H. & Visaltanachoti, N. (2017) Time series momentum and moving average trading rules. *Quantitative Finance*, 17(3), 405–421.

Martin-Valmayor, M.A., Gil-Alana, L.A. & Infante, J. (2023) Energy prices in europe. evidence of persistence across markets. *Resources Policy*, 82, 103546.

Maryniak, P. & Weron, R. (2019) What is the probability of an electricity price spike? evidence from the uk power market. In: *Handbook of Energy Finance: Theories, Practices and Simulations*World Scientific, pp. 231–245.

Mitrentsis, G. & Lens, H. (2022) An interpretable probabilistic model for short-term solar power forecasting using natural gradient boosting. *Applied Energy*, 309, 118473.

Molnar, C. (2020) *Interpretable machine learning*. : Lulu. com.

Niazkar, M., Menapace, A., Brentan, B., Piraei, R., Jimenez, D., Dhawan, P. et al. (2024) Applications of xgboost in water resources engineering: A systematic literature review (dec 2018–may 2023). *Environmental Modelling & Software*,, 105971.

Nowotarski, J. & Weron, R. (2018) Recent advances in electricity price forecasting: A review of probabilistic forecasting. *Renewable and Sustainable Energy Reviews*, 81, 1548–1568.

Packham, N., Papenbrock, J., Schwendner, P. & Woebbeking, F. (2017) Tail-risk protection trading strategies. *Quantitative Finance*, 17(5), 729–744.

Rawson, A., Brito, M. & Sabeur, Z. (2022) Spatial modeling of maritime risk using machine learning. *Risk Analysis*, 42(10), 2291–2311.

Refinitiv (2024) *Refinitiv*. Accessed: 2024-05-23.
URL https://www.lseg.com/en

Ruisen, L., Songyi, D., Chen, W., Peng, C., Zuodong, T., YanMei, Y. et al. Bagging of xgboost classifiers with random under-sampling and tomek link for noisy label-imbalanced data. In: *IOP Conference Series: Materials Science and Engineering*. Vol. 428. *IOP Publishing, 2018*, p. 012004.

Saâdaoui, F. & Jabeur, S.B. (2023) Analyzing the influence of geopolitical risks on european power prices using a multiresolution causal neural network. *Energy Economics*, 124, 106793.

Sandhu, H.S., Fang, L. & Guan, L. (2016) Forecasting day-ahead price spikes for the ontario electricity market. *Electric Power Systems Research*, 141, 450–459.

Shapley, L.S. et al. (1953) A value for n-person games,.

Sokolova, M., Japkowicz, N. & Szpakowicz, S. Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. In: *Australasian Joint Conference on Artificial Intelligence*. *Springer, 2006*, pp. 1015–1021.

Souhir, B.A., Heni, B. & Lotfi, B. (2019) Price risk and hedging strategies in nord pool electricity market evidence with sector indexes. *Energy Economics*, 80, 635–655.

Srinivas, P. & Katarya, R. (2022) hyoptxg: Optuna hyper-parameter optimization framework for predicting cardiovascular disease using xgboost. *Biomedical Signal Processing and Control*, 73, 103456.

Stathakis, E., Papadimitriou, T. & Gogas, P. (2021) Forecasting price spikes in electricity markets. *Review of Economic Analysis*, 13(1), 65–87.

Stuke, A., Rinke, P. & Todorović, M. (2021) Efficient hyperparameter tuning for kernel ridge regression with bayesian optimization. *Machine Learning: Science and Technology*, 2(3), 035022.

Su, C.W., Khan, K., Umar, M. & Zhang, W. (2021) Does renewable energy redefine geopolitical risks? *Energy Policy*, 158, 112566.

Tafakori, L., Pourkhanali, A. & Fard, F.A. (2018) Forecasting spikes in electricity return innovations. *Energy*, 150, 508–526.

Tanha, J., Abdi, Y., Samadi, N., Razzaghi, N. & Asadpour, M. (2020) Boosting methods for multi-class imbalanced data classification: an experimental review. *Journal of Big Data*, 7, 1–47.

Tjoa, E. & Guan, C. (2020) A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Transactions on Neural Networks and Learning Systems*, 32(11), 4793–4813.

Trebbien, J., Gorjão, L.R., Praktiknjo, A., Schäfer, B. & Witthaut, D. (2023) Understanding electricity prices beyond the merit order principle using explainable ai. *Energy and AI*, 13, 100250.

Trueck, S., Weron, R. & Wolff, R. (2007) Outlier treatment and robust approaches for modeling electricity spot prices,.

Tschora, L., Pierre, E., Plantevit, M. & Robardet, C. (2022) Electricity price forecasting on the day-ahead market using machine learning. *Applied Energy*, 313, 118752.

Turner, R., Eriksson, D., McCourt, M., Kiili, J., Laaksonen, E., Xu, Z. et al. Bayesian optimization is superior to random search for machine learning hyperparameter tuning: Analysis of the black-box optimization challenge 2020. In: *NeurIPS 2020 Competition and Demonstration Track*. *PMLR, 2021*, pp. 3–26.

Ullah, M.H., Paul, S. & Park, J.D. Real-time electricity price forecasting for energy management in grid-tied mtdc microgrids. In: *2018 IEEE Energy Conversion Congress and Exposition (ECCE). IEEE, 2018*, pp. 73–80.

Varmedja, D., Karanovic, M., Sladojevic, S., Arsenovic, M. & Anderla, A. Credit card fraud detection-machine learning methods. In: *2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH). IEEE, 2019*, pp. 1–5.

Wang, C., Deng, C. & Wang, S. (2020) Imbalance-xgboost: leveraging weighted and focal losses for binary label-imbalanced classification with xgboost. *Pattern Recognition Letters*, 136, 190–197.

Westgaard, S., Fleten, S.E., Negash, A., Botterud, A., Bogaard, K. & Verling, T.H. (2021) Performing price scenario analysis and stress testing using quantile regression: A case study of the californian electricity market. *Energy*, 214, 118796.

Yamada, K. & Mori, H. A deep learning technique for electricity price forecasting in consideration of spikes. In: *TENCON 2021-2021 IEEE Region 10 Conference (TENCON). IEEE, 2021*, pp. 744–749.

Yang, L. & Shami, A. (2020) On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, 415, 295–316.

Yao, S., Wu, Q., Kang, Q., Chen, Y.W. & Lu, Y. (2024) An interpretable xgboost-based approach for arctic navigation risk assessment. *Risk Analysis*, 44(2), 459–476.

Zamudio López, M., Zareipour, H. & Quashie, M. (2024) Forecasting the occurrence of electricity price spikes: A statistical-economic investigation study. *Forecasting*, 6(1), 115–137.

Zhang, C., Tan, K.C., Li, H. & Hong, G.S. (2018) A cost-sensitive deep belief network for imbalanced classification. *IEEE Transactions on Neural Networks and Learning Systems*, 30(1), 109–122.

Zhang, J., Ma, X., Zhang, J., Sun, D., Zhou, X., Mi, C. et al. (2023) Insights into geospatial heterogeneity of landslide susceptibility based on the shap-xgboost model. *Journal of Environmental Management*, 332, 117357.

Zhang, Z., He, M., Zhang, Y. & Wang, Y. (2022) Geopolitical risk trends and crude oil price predictability. *Energy*, 258, 124824.

Zou, Q., Xie, S., Lin, Z., Wu, M. & Ju, Y. (2016) Finding the best classification threshold in imbalanced classification. *Big Data Research*, 5, 2–8.

MANCHESTER 1824
The University of Manchester

An Improved Exact Algorithm for the Electric Vehicle Routing Problem in V2G Energy Network

Date of submission
2025

Yutong Qi

Management Sciences and Marketing

List of revisions

Contents

Word count: ???

1 Abstract

The integration of electric vehicles (EVs) into smart grids through vehicle-to-grid (V2G) technology introduces complex challenges in coordinating mobility and energy management. This study addresses the Vehicle-to-Grid Electric Vehicle Routing Problem with Time Windows (V2G-EVRPTW), which optimizes EV fleet routing and bidirectional energy transactions under time-of-use (TOU) electricity pricing. The objective is to minimize net electricity costs by strategically scheduling charging during low-price periods and discharging during high-price periods, while ensuring customer time-window constraints and battery capacity limits are met. We formulate V2G-EVRPTW as a mixed-integer linear program and propose an exact branch-and-price algorithm enhanced with a tailored column generation framework. Our approach incorporates a K-best column retrieval strategy with adaptive column count adjustment, significantly improving convergence speed. Computational experiments on adapted EVRPTW instances demonstrate that our method outperforms a Gurobi baseline and a standard branch-and-price variant, achieving faster bound tightening and superior solution quality within fixed time limits. These results highlight the algorithm's efficacy in balancing logistics efficiency with grid-oriented energy optimization.

2 Introduction

[MANUEL: I fixed the compilation errors shown in Overleaf. Please do not give us a document that has compilation errors.]

[MANUEL: For the thesis, the citations should be author-year. Please ask Ozioma Paul to share the .cls file and the tempalte she used for her thesis. I think you are using a different template.]

With the growing awareness of environmental protection and sustainable development, electric vehicles are gradually emerging as an essential means of transportation in the logistics and distribution sectors. However, compared with traditional internal combustion engine vehicles, electric vehicles inherently face limitations in driving range, charging time, and energy management[1].[1][2] These challenges necessitate targeted modifications and extensions to the conventional Vehicle Routing Problem (VRP). To address these issues, the Electric Vehicle Routing Problem with Time Windows (EVRPTW) has been introduced. This problem not only aims to optimize transportation costs under the constraints of customer demands and time windows but also requires careful consideration of key factors such as battery capacity, charging strategies, and energy consumption models.

Vehicle-to-grid (V2G) connectivity is a promising concept, it remains in the pilot development stage. In the medium to long term, the rapid proliferation of electric vehicles (EVs) is expected to exert significant pressure on power distribution systems[2]. On one hand, EV charging infrastructure is anticipated to account for a substantial share of total electricity demand, potentially undermining

[1] **[MANUEL: Before cite commands there should be a space.]**

[2] **[MANUEL: If you have the doi you do not need to provide the URL. And you do not need to provide the "visited" for journal/conference papers]**

grid stability and operational efficiency. Indeed, uncoordinated charging of a large EV fleet may lead to severe power losses, voltage deviations, and substation overloads[3].

Under unidirectional G2V scenarios (i.e., grid-to-vehicle), these challenges can be partially mitigated by controlling charging schedules or utilizing distributed energy storage, where vehicles are charged during recommended off-peak intervals[4]. On the other hand, in bidirectional V2G settings (i.e., vehicle-to-grid), smart grids enable two-way energy exchange, allowing EVs to inject electricity back into the grid. When properly coordinated, such interactions can support peak shaving and ancillary services[5]. By jointly optimizing charging and discharging behaviors, EV fleets can help flatten demand curves and enhance the integration of intermittent renewable energy sources.

The scheduling of EV operations within smart grid environments is inherently sensitive to electricity pricing schemes. Rasheed et al.[6][3] proposed a distributed pricing mechanism, while [7][4] introduced a coordinated dynamic pricing framework. Xu et al.[8][5] further developed a zonal and time-based electricity tariff policy. These pricing mechanisms aim to dynamically adjust incentives to shift EV charging to off-peak periods, thereby improving both economic and grid performance.

3 Literature review

3.1 Electric Vehicle Routing Problems

The major problem electric vehicles (EVs) face is battery limitation by delivery[9]. Grunditz and Thiringer[10] conducted an analysis of more than 40 electric vehicles (EVs) available worldwide, classifying them into small, medium-to-large, high-performance, and sports categories. All of these EVs have battery capacities ranging from 12 to 90 kWh and driving distances varying between 85 and 528 km. For delivery services, light vans and freight EVs are primarily utilized, though they generally offer a shorter range (160-240 km) compared to internal combustion engine vehicles (ICEVs), which typically have a range of 480-650 km. Thus, for electric vehicle routing problems (EVRP), energy consumption and charge problem should be highlighted.

3.1.1 Energy Consumption

In the current research, longitudinal dynamics modeling (LDM) is frequently used to accurately estimate energy consumption. Firstly, the underlying dynamics model is introduced by Asamer et al[11]. The force F) is defined by equation(1.1) Here, m represents the vehicle's mass (typically when empty), a stands for acceleration, v is the vehicle speed, g denotes the gravitational constant, and f accounts for the inertia force of the vehicle's rotating parts. Additionally, α indicates the road slope, c_r is the rolling friction coefficient, c_d is the air drag coefficient, ρ refers to air density, and A is the frontal surface area of the vehicle.

[3] **[MANUEL: Use Rasheed et al. [6], do not write authors' names yourself]**
[4] **[MANUEL: same]**
[5] **[MANUEL: same]**

$$F = mg\sin\alpha + c_r mg\cos\alpha + 0.5 c_d \rho A v^2 + fma \tag{1.1}$$

And the energy consumption P_b can be calculated by (1.2). Where μ_m represents the transmission coefficient between the electric motor and drivetrain, μ_e is the conversion ratio from chemical to electric energy, and μ_g is the ratio for converting mechanical energy back to chemical energy in the battery. Energy is returned to the battery only when the force is negative and the speed exceeds the minimum limitation [11].[6]

$$P_b = \begin{cases} \mu_e(\mu_m F_v + P_0), & \text{if } F \geq 0 \\ \begin{cases} 0, & \text{if } v \leq v_{\text{min}} \\ \mu_g F_v + P_0, & \text{else} \end{cases}, & \text{if } F < 0 \end{cases} \tag{1.2}$$

[7]

Goeke and Schneider [12] enhanced the energy consumption model by incorporating changes in vehicle load mass during goods transport, excluding acceleration and deceleration phases, thereby refining the solution's accuracy.
In reality, cars travel at variable speeds, making it challenging to consistently determine acceleration and braking. For electric vehicle routing problems, researchers often simplify the longitudinal dynamics modeling (LDM) to accommodate these irregularities. Lera-Romero[13] developed a generic framework for time-dependent Electric Vehicle Routing Problems (EVRP), where battery consumption P[8] is influenced by variables grouped into three categories: vehicle mass, speed, and terrain conditions, as shown in equation (1.3). Where m_c and m_q denote kerb mass and goods mass separately.

$$P(v,q) = \left(\frac{1}{2}\cdot c_d \cdot \rho \cdot A \cdot v^2 + (m_c + m_q)\cdot g \cdot (\sin(\alpha) + c_r \cdot \cos(\alpha))\right)\cdot v \tag{1.3}$$

Zhang [14] applies a similar method to calculate mechanical power and then estimates the indirect carbon dioxide emissions associated with generating the battery energy, based on the quantity of battery energy consumed.

3.1.2 Charging Problem

The topic of charging electric vehicles can be divided into three discussion areas: charging or swapping batteries, various charging measurement rates, and the features of charging stations.
Jie and Yang [15] adapted the two-echelon Electric Vehicle Routing Problem (2E-EVRP) by replacing charging stations with battery switching stations. They employed a hybrid algorithm that combines column generation and adaptive large neighborhood search (CG-ALNS) to optimize compu-

[6] **[MANUEL: This is not the standard way to write piece-wise functions, write the conditions as if $F < 0 \ \wedge \ v \leq v_{\text{min}}$]**
[7] **[MANUEL: To star a new paragraph just type an empty line, you do not need \\]**
[8] **[MANUEL: space before math symbols]**

tations for these battery switching stations. Raeesi [16] introduced an advanced Electric Vehicle Routing Problem (EVRP) that incorporates time windowing and simultaneous mobile battery swapping. This approach transitions the swapping process from stationary stations to mobile vans, allowing both the exchange vans and distribution vehicles to move simultaneously to optimize the planning of the path, thus reducing time costs. Batteries' inherent properties impose a typical limit of 80 percent of their capacity for charging [17]. Up to this limit, charging can proceed at a fast linear rate, but any charging beyond this threshold must occur at a slower, non-linear rate to prevent battery damage from overcharging. Montoya et al. [18] introduced a realistic model of vehicle charging using a nonlinear function, which they named the Electric Vehicle Path Problem with Nonlinear Charging (E-VRP-NL). This model takes into account various charging technologies and categorizes charging stations into three types: slow, medium, and rapid, each associated with specific linear and nonlinear charging rates. The study highlighted that neglecting the nonlinear aspects of charging can result in unfeasible or excessively costly solutions.
However, Wang[19] expanded the Electric Fleet Size and Mix Vehicle Routing Problem with Time Windows and Recharging Stations (E-FSMFTW) by incorporating partial linear charging strategies. This modification revealed that allowing partial linear charging significantly lowers the logistic costs for large E-FSMFTW instances when compared to the best-known solutions that rely solely on full charging strategies.

3.2 Vehicle Routing Problem Algorithm

Since the VRP and its variants are essentially NP-hard problems that cannot obtain an exact solution in a short time, many related algorithms have been derived to simplify the computational steps. Anuar [20] classifies the algorithms of VRP into three categories, namely, exact, heuristic, and meta-heuristic (see Figure 1).[9].

3.2.1 Approximate Methods

Heuristic algorithms are classified into three main categories, i.e., constructive, two-stage and local improvement heuristics. Gábor Nagy and Saïd Salhi[21] developed an integrated heuristic approach that does not rely on the assumption prevalent in the traditional VRPPD literature that goods can only be picked up after all deliveries have been made. Their proposed approach integrates the pickup and delivery problems by modifying heuristic routines extracted from VRP methods to reduce infeasible solutions while constructing problem-specific routines. Many literatures focus on metaheuristic algorithms, which are more advanced procedures that incorporate population search and local search methods. In the domain of home healthcare logistics, Liu [22] proposed two mixed integer planning models and developed Genetic Algorithm (GA) and Taboo Search (TS) methods considering the special vehicle scheduling problems of simultaneous delivery and pickup and time windows.GA is based on aligning chromosomes, segmentation process and local search, while TS is based on route assignment attributes, augmented cost function, route re-optimization and route

[9] **[MANUEL: Use \ref and \label commands not explicit numbers: Figure 1]**

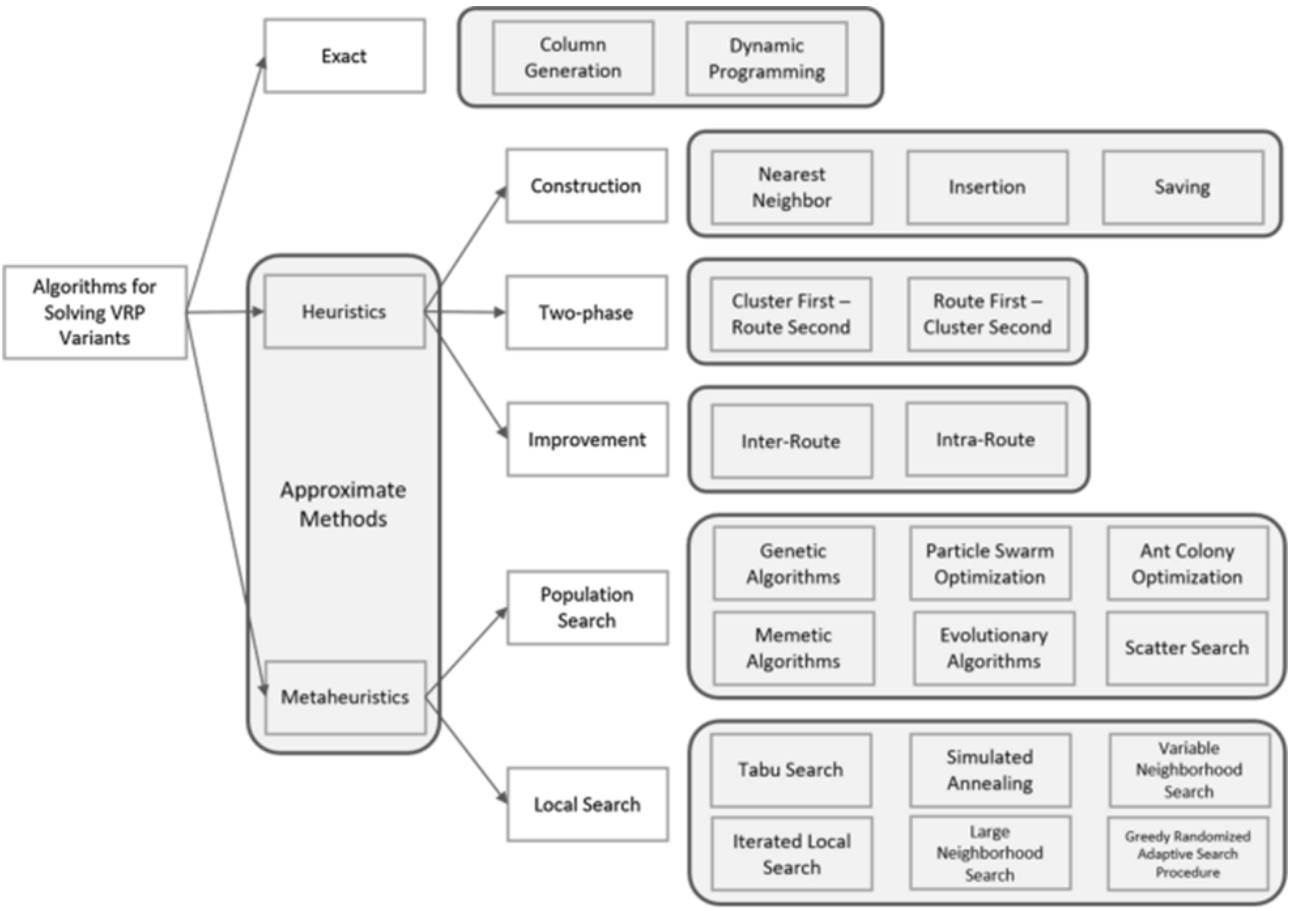


Fig. 1. Classification of VRP algorithms**[MANUEL: This is blurry. Export it from the original as PDF (converting from PNG to PDF will not be enough)]**

optimization based on the attribute desire level. These heuristic algorithms were tested on an existing Vehicle Routing Problem with Time Window (VRPTW) benchmarking example and showed effectiveness and superiority in solving home healthcare logistics problems. For the VRP with Simultaneous Pickup and Delivery problem, Öztaş and Tuş [23] have designed a hybrid meta-heuristic algorithm that combines meta-heuristic techniques such as iterative local search (ILS), Variable Neighbourhood Descent (VND), and Threshold Acceptance (TA) in order to search for a near-optimal solution in a reasonable amount of time. By using a roulette wheel selection mechanism based on Shannon entropy and an adaptive acceptance criterion, the algorithm efficiently explores and enhances the search space to improve path feasibility and reduce total cost.

3.2.2 Exact Algorithms

The exact algorithms mainly include Lagrangian relaxation and column generation. Tang [24] proposed an innovative solution that cleverly integrates the Lagrangian relaxation technique with the reinforcement learning framework. With this approach, the investigators were able to transform complex soft constraint problems into more manageable multi-objective optimisation problems. In their work, the Lagrangian relaxation not only effectively handles the trade-off between travelling distance and constraint violation cost, but also improves the efficiency and effectiveness of the algorithm in finding near-optimal solutions. Zang[25] uses column generation as the core technique to break the complex EVRP into smaller, more manageable subproblems. They carefully design

custom labeling algorithms and resource extension functions that generate efficient columns, enabling the model to effectively balance nonlinear battery depreciation with travel costs. This approach leads to faster convergence toward near-optimal solutions. Caceres[26] centers on a column generation strategy to simplify the challenging school bus routing problem for special education students. By combining a greedy heuristic with a column generation framework, the authors decompose the problem into tractable subproblems that address mixed loading and heterogeneous fleet configurations. The method iteratively identifies effective routes, ultimately reducing fleet size and overall travel cost while ensuring high service quality.

3.3 Vehicle to Grid Optimization

Vehicle-to-grid (V2G) interaction has been widely discussed in the energy systems literature, yet its integration with time-varying electricity prices and vehicle routing optimization remains relatively underexplored in both academia and industry. Yao et al. [27] investigated the impact of spatiotemporal electricity pricing on routing decisions by proposing a bi-level framework in which fleet operators optimize routes and charging schedules to minimize costs, while offering monetary incentives to compensate customers for delivery time flexibility. Customers, in turn, adapt their time windows based on cost trade-offs. Barco et al. [28] incorporated a detailed energy consumption model using the Lightweight Dynamic Model (LDM), and explored routing optimization under dynamic electricity tariffs. However, their work did not consider vehicle-to-grid discharging or energy trading decisions.

In bidirectional V2G environments, electric vehicles can inject energy back into the grid. Tang et al. [29] proposed a smart grid-based EV network that incorporates heterogeneous charging facilities with varying costs and capacities. Customers select routes and determine when and where to charge or discharge in order to minimize total cost. Triviño et al. [30] extended this idea by proposing a joint charging/discharging and routing strategy for operating EV fleets, though their model assumes fixed delivery routes and focuses primarily on the economic gains from energy trading, rather than demand fulfillment or route optimization.

Building on this, Liu et al. [31] introduced additional complexities such as intermittent renewable generation, limited charging infrastructure, and EV delay tolerance. To reduce computational complexity, their approach decouples routing and energy scheduling: routes are determined via A* search, followed by charging/discharging decisions. However, this separation may lead to suboptimal solutions when electricity price variation is high, as profitable detours for energy transactions may be neglected.

Lin et al. [32] further advanced this line of work by developing a multi-period routing model that leverages EV battery storage capacity to flexibly manage charging and discharging, aiming to maximize fleet-level profit. Their heuristic approach was evaluated on a grocery delivery EV fleet operating in the Kitchener–Waterloo region of Ontario, Canada. Nonetheless, their model imposes a restriction that all charging decisions must occur at the beginning of each period, which limits temporal flexibility and responsiveness to real-time price fluctuations.

In contrast, the present study addresses this research gap by proposing a generalizable V2G-EVRP model that integrates time-dependent electricity pricing and bidirectional energy flow into vehicle

Table 1. Existing literature on EV routing and V2G operations.

	Customer satisfaction	Discharging	Time-variant prices	Flexible charging	Approach
Schneider et al. (2014)	✓			✓	Exact
Yao et al. (2023)	✓		✓	✓	Heuristic
Barco et al. (2017)	✓		✓	✓	Exact
Tang et al. (2017)	✓	✓	✓	✓	Exact
Triviño-Cabrera et al. (2019)	✓	✓	✓		Heuristic
Liu et al. (2020)	✓	✓	✓		Exact
Lin et al. (2021)	✓	✓	✓		Heuristic
V2G-EVRP (this work)	✓	✓	✓	✓	Exact

routing decisions. Building on the work of Liu and Lin, we allow charging and discharging decisions to be made flexibly at any time during each period, depending on dynamic electricity prices. Moreover, the model is solved using an exact algorithm, providing high-quality benchmark solutions to facilitate future comparisons and evaluation of heuristic approaches.

4 Problem Statement and Mathematical Formulation

4.1 Problem description (Framework and assumptions)

We investigate the Vehicle-to-Grid Electric Vehicle Routing Problem with Time Windows under time-of-use electricity pricing (V2G-EVRPTW-TOU). The problem considers a road network with customer locations, charging stations, and depot location.

A homogeneous fleet of electric vehicles starts from the origin depot with fully charged batteries and must visit all customers to fulfill their delivery demands before returning to the destination depot within a given planning horizon. Each customer must be served within a specific time window, and service requires a certain amount of time at the customer location. Vehicles consume energy for traveling between locations and may need to recharge at charging stations or depot to maintain sufficient battery levels. In addition to charging, vehicles can also discharge energy back to the grid (Vehicle-to-Grid service) at charging stations and depot. Electricity prices vary throughout the day according to a time-of-use tariff structure.

At charging stations and depots, vehicles have flexibility to choose when to start charging or discharging and how long to perform each action. They can wait if needed to align these energy transactions with periods of favorable electricity prices. The actual energy exchanged depends on the timing and duration of these actions relative to the changing price periods.

The objective is to determine vehicle routes, customer service times, and energy management strategies that minimize total net electricity cost (purchases minus sales), while satisfying all customer service, capacity, time window, and battery constraints. Upon arriving at the destination depot, vehicles must recharge to full capacity for the next day's operations with complete timing flexibility; the associated cost is included in the objective based on time-of-use prices.

4.2 Math Notation

The V2G-EVRPTW under TOU pricing is defined on directed network $G = (V, A)$. The node set V consists of customer nodes $C = \{1, \ldots, n\}$, charging station nodes $S = \{n+1, \ldots, n+s\}$, origin depot 0, and destination depot $d = n + s + 1$. Each customer $i \in C$ requires delivery of demand $q_i \leq C^{\text{veh}}$ with service beginning in time window $[a_i, b_i]$.

Each arc $(i, j) \in A$ is characterized by travel time τ_{ij} (incorporating service time at i), distance d_{ij}, and traction energy consumption $e_{ij} = g \cdot d_{ij}$ where $g > 0$ is the constant energy-per-distance rate (kWh/km).

A homogeneous fleet of $K = \{1, \ldots, K_{\text{max}}\}$ electric vehicles, each equipped with battery capacity Q and payload capacity C^{veh}, departs from depot 0 with full charge $B_{0k} = Q$ during time window $[a_0, b_0]$. The planning horizon spans $[0, H]$. The battery state of vehicle $k \in K$ at node $i \in V$ is tracked by arrival state-of-charge $B_{ik} \in [0, Q]$ and departure state-of-charge $B_{ik}^{\text{dep}} \in [0, Q]$.

Charging and discharging are permitted at energy nodes $E = S \cup \{0, d\}$. TOU tariffs are modeled by discretizing the horizon $[0, H]$ into $|T|$ equal-length periods $t \in T$ of duration $\delta = H/|T|$. At energy nodes $i \in E$, vehicle k determines the start time $\sigma_{ikt} \geq 0$ and duration $\ell_{ikt} \geq 0$ of each energy transaction in period t, subject to maximum power rates $P^+,\ P^-$.

The effective energy exchanged in period t is calculated as $E_{ikt} = P^+ \cdot \min\{\ell_{ikt}, \delta - \sigma_{ikt}\}^+$ for charging ($E_{ikt} > 0$) or $E_{ikt} = -P^- \cdot \min\{\ell_{ikt}, \delta - \sigma_{ikt}\}^+$ for discharging ($E_{ikt} < 0$). Vehicle-to-Grid (V2G) revenue modeling credits vehicles with revenue $-p_t^- E_{ikt}$ for each unit of energy discharged ($E_{ikt} < 0$), which directly offsets the charging costs $p_t^+ E_{ikt}$ incurred during charging ($E_{ikt} > 0$). This bidirectional energy flow enables vehicles to generate profit by selling excess battery capacity to the grid during low-demand, high-price periods.

The battery state evolves according to:

$$B_{ik}^{\text{dep}} = B_{ik} + \sum_{t \in T} E_{ikt} - e_{ij}, \quad \forall (i,j) \in A, k \in K,$$

bounded by $0 \leq B_{ik}, B_{ik}^{\text{dep}} \leq Q$. At depot d, vehicles must achieve $B_{dk}^{\text{dep}} = Q$ to ensure full readiness for next-day operations.

Battery feasibility constraints guarantee safe operation:

- *Arrival safety:* $B_{jk} \geq e_{ij}$ if $x_{ijk} = 1$ (sufficient charge to complete trip to j)
- *No overcharge:* $B_{ik}^{\text{dep}} \leq Q$ after charging
- *No deep discharge:* $B_{ik}^{\text{dep}} \geq 0$ after discharging
- *Continuous tracking:* $B_{jk} = B_{ik}^{\text{dep}} - e_{ij}$ if $x_{ijk} = 1$

The objective minimizes total, calculated as $\sum_{i \in E} \sum_{k \in K} \sum_{t \in T} c_{ikt} E_{ikt}$, where $c_{ikt} = p_t^+$ if $E_{ikt} > 0$ and $c_{ikt} = -p_t^-$ if $E_{ikt} < 0$, subject to customer service, vehicle capacity $\sum_{i \in C} q_i x_{ijk} \leq C^{\text{veh}}$, time-window, and battery constraints.

Parameters and decision variables appear in Tables 2 and 3.

Symbol	Meaning
Index/identifier parameters	
$0,\ d = n + s + 1$	Indices of origin and destination depots
$n,\ s$	Number of customers and charging stations ($C = \{1, \ldots, n\}$, $S = \{n+1, \ldots, n+s\}$)
Network, time, and demand	
$G = (V, A)$	Directed network; $V = \{0\} \cup C \cup S \cup \{d\}$, $A \subseteq V \times V$
$[a_i, b_i]$	Time window at node $i \in C \cup \{0, d\}$
q_i	Demand at customer $i \in C$
H	Planning horizon length
Travel and traction energy	
τ_{ij}	Travel time on arc $(i,j) \in A$ (includes service time at i)
d_{ij}	Distance on arc $(i,j) \in A$
e_{ij}	Traction energy on arc $(i,j) \in A$; $e_{ij} = g\, d_{ij}$
$g > 0$	Constant energy-per-distance coefficient
Vehicle, battery, and capacity	
$K_{\max}$	Upper bound on number of vehicles ($K = \{1, \ldots, K_{\max}\}$)
Q	Battery capacity
C^{veh}	Vehicle payload capacity
Time periods and pricing	
$T,\ \delta = H/\lvert T\rvert$	Set of time periods and period length
$p_t^+,\ p_t^-$	Buy (charging)/sell (discharging) price in period $t \in T$
Power limits	
$P^+,\ P^-$	Maximum charging/discharging power
M_T	A Large constant for time constraint relaxation
M_B	A Large constant for battery state constraint relaxation

Table 2. Parameters

Symbol	Meaning
Binary variables	
x_{ijk}	1 if vehicle k traverses arc (i,j), 0 otherwise
z_{ik}^+	1 if vehicle k charges at node i, 0 otherwise
z_{ik}^-	1 if vehicle k discharges at node i, 0 otherwise
Continuous variables	
T_{ik}	Arrival time of vehicle k at node i
W_{ik}	Dwell time of vehicle k at node i
B_{ik}	Arrival state-of-charge of vehicle k at node i
B_{ik}^{dep}	Departure state-of-charge of vehicle k at node i
σ_{ikt}	Start time of energy transaction of k at i in period t
ℓ_{ikt}	Duration of energy transaction of k at i in period t
E_{ikt}	Net energy exchanged by k at i in period t (> 0: charging)

Table 3. Decision variables

4.3 Mathematical Model and Explanation

All energy transactions at energy nodes $E = S \cup \{0, d\}$ (including destination depot d) are priced according to period-wise TOU tariffs $p_t^+,\ p_t^-$. The cost of final charging to full capacity at d is captured by energy exchanges E_{dkt} during periods $t \in T$. Vehicles must depart d at full charge $B_{dk}^{\text{dep}} = Q$ for next-day operations.

Objective (net electricity cost).

$$\text{Minimize} \underbrace{\sum_{i \in S \cup \{0\}} \sum_{k \in K} \sum_{t \in T} \left(p_t^+ P^+ \ell_{ikt} - p_t^- P^- \ell_{ikt} \right)}_{\text{net charging/discharging cost at stations and origin}} + \underbrace{\sum_{k \in K} p_t^+ \left(Q - B_{dk} \right)}_{\text{“fill-to-full” cost: charge } (Q - B_{dk}) \text{ at destination}} . \tag{2.1}$$

Routing constraints:

$$\sum_{k \in K} \sum_{j:(i,j) \in A} x_{ijk} = 1 \qquad \forall i \in C \tag{2.2}$$

$$\sum_{j:(0,j) \in A} x_{0jk} \le 1, \qquad \sum_{i:(i,d) \in A} x_{idk} \le 1 \qquad \forall k \in K \tag{2.3}$$

$$\sum_{j:(i,j) \in A} x_{ijk} = \sum_{h:(h,i) \in A} x_{hik} \qquad \forall i \in V \setminus \{0, d\},\ k \in K \tag{2.4}$$

Time windows and propagation.

$$T_{jk} \ge T_{ik} + \tau_{ij} - M_T (1 - x_{ijk}) \qquad \forall (i,j) \in A,\ k \in K \tag{2.5}$$

$$a_i \le T_{ik} \le b_i \qquad \forall i \in C \cup \{0, d\},\ k \in K \tag{2.6}$$

$$\sum_{t \in T} \ell_{ikt} \le W_{ik} \qquad \forall i \in E,\ k \in K \tag{2.7}$$

$$\ell_{ikt} \le \delta - \sigma_{ikt} + M_T (1 - z_{ik}^+) \qquad \forall i \in E,\ t \in T,\ k \in K \tag{2.8}$$

Power bounds and mode exclusivity. Charging stations may be visited without energy transactions ($z_{ik}^+ + z_{ik}^- = 0$ allowed for waiting/time alignment). Both constraints are implicit via E_{ikt} bounds:

$$0 \le \ell_{ikt} \le \bar{\ell}\, z_{ik}^+, \qquad 0 \le \ell_{ikt} \le \bar{\ell}\, z_{ik}^-, \qquad z_{ik}^+ + z_{ik}^- \le 1 \qquad \forall i \in E,\ t \in T,\ k \in K \tag{2.9}$$

Energy balance and capacity. Simultaneous charging/discharging in different periods is permitted for TOU arbitrage.

$$B_{ik}^{\text{dep}} = \begin{cases} B_{ik} + \sum_{t \in T} E_{ikt}, & i \in E, \\ B_{ik}, & i \in C, \end{cases} \qquad \forall i \in V,\ k \in K \tag{2.10}$$

$$B_{jk} \geq B_{ik}^{\mathsf{dep}} - e_{ij} - M_B(1 - x_{ijk}) \qquad \forall (i,j) \in A,\, k \in K \tag{2.11}$$

$$0 \leq B_{ik} \leq Q, \qquad 0 \leq B_{ik}^{\mathsf{dep}} \leq Q \qquad \forall i \in V,\, k \in K \tag{2.12}$$

Terminal requirements.

$$\sum_{j:(0,j)\in A} x_{0jk} = \sum_{i:(i,d)\in A} x_{idk}, \qquad B_{0k} = Q, \qquad B_{dk}^{\mathsf{dep}} = Q \qquad \forall k \in K \tag{2.13}$$

Variable domains.

$$x_{ijk}, z_{ik}^{+}, z_{ik}^{-} \in \{0,1\}; \quad T_{ik}, W_{ik}, \sigma_{ikt}, \ell_{ikt}, E_{ikt}, B_{ik}, B_{ik}^{\mathsf{dep}} \geq 0 \tag{2.14}$$

Equation (2.1) decomposes the total cost into two transparent parts: the net charging/discharging cost accrued at stations and the origin depot, and the *fill-to-full* cost at the destination. TOU pricing enters via period-wise buy/sell prices $\{p_t^+, p_t^-\}$ and linear "time $\times$ power" accounting: in period t, the effective charged energy is $P^+\ell_{ikt}$ and the effective discharged energy is $P^-\ell_{ikt}$. At the destination d, the charging overlaps $\{\ell_{dkt}\}$ priced by $\{p_t^+\}$ naturally capture the cost of filling after arrival. We do not prohibit discharging at d followed by charging (e.g., arbitrage), but such behavior must respect time-window feasibility and the single-mode exclusivity below and must culminate in leaving d at full charge through (2.13).

Equations (2.2)–(2.4) form a standard arc-flow routing structure: (2.2) enforces that each customer is visited exactly once; (2.3) limits each vehicle to at most one departure from the origin and one return to the destination; (2.4) maintains flow conservation at all intermediate nodes so that each vehicle's path continuously connects 0 to d. Temporal feasibility is enforced by (2.5)–(2.8). Equation (2.5) propagates arrival times along arcs and ties them to arc selection through the big-M_T term; (2.6) enforces time windows at customers and at the depots; (2.7) caps the total charging/discharging overlap per node by the node dwell time; (2.8) requires every positive overlap to lie within the effective availability window of the node (or its price segment), thereby ensuring that transactions occur only during valid periods and are consistently priced.

Physical consistency of single-mode operation within any node-period is enforced by (2.9): we link the overlap durations to binary mode indicators via a unified upper bound $\bar{\ell}$, and impose $z_{ik}^+ + z_{ik}^- \leq 1$ so that charging and discharging cannot occur simultaneously at the same node and period. Energy dynamics and capacity are captured by (2.10)–(2.12). At energy nodes, the departure SoC equals the arrival SoC plus the efficiency-adjusted charging input minus the efficiency-adjusted discharging output; along arcs, the arrival SoC must exceed the departure SoC minus traction energy (with big-M_B to deactivate unused arcs); all arrival and departure SoC variables lie within $[0, Q]$. Finally, (2.13) enforces departure/return consistency per vehicle, full-charge release at the origin, and *fill-to-full at the destination after arrival* (i.e., we require $B_{dk}^{\mathsf{dep}} = Q$ rather than full-charge *upon* arrival). Together with (2.1), this yields a rigorous realization of "minimize cost $=$ net charging/discharging $+$ fill-to-full" under TOU pricing. The model remains entirely linear in pricing, time

windows, power limits, and efficiencies, thereby integrating seamlessly with a column-generation / branch-and-price framework.

5 Algorithm Framework

5.1 Column Generation

(need cite) The Vehicle Routing Problem (VRP) and its variants, including the Electric Vehicle Routing Problem (EVRP), are large-scale combinatorial optimization problems where the exponential number of possible routes renders straightforward set-partitioning formulations computationally intractable for realistic instance sizes.

Column generation (CG) addresses this challenge through a decomposition technique that iteratively solves a Restricted Master Problem (RMP) and a Pricing Subproblem. By generating only promising routes with negative reduced cost on demand, CG avoids enumerating all possibilities upfront. This separation allows the RMP to focus solely on global coupling constraints (e.g., customer coverage), while the Pricing Subproblem handles complex within-route resources (e.g., time windows, SOC propagation) via dynamic programming or labeling algorithms. Consequently, CG produces tight LP bounds and scales efficiently, making it ideal for variants with added constraints like energy management in EVRP. Baldacci et al. (2011) established CG as the cornerstone of state-of-the-art branch-and-price frameworks for the Capacitated VRP, while Afsar et al. (2014) demonstrated its flexibility for generalized VRPs with variable fleet sizes. In EVRP contexts, Kozák et al. (2020) adapted CG-like route generation to handle energy constraints, confirming the framework's applicability despite increased complexity.

Restricted Master Problem (RMP)

Let P denote the set of *all* time-and-energy-feasible *elementary* routes starting at 0 and ending at d, where an elementary route visits a subset of customers exactly once each and may include charging stations S for energy replenishment/discharge. Thus, $|P| = n$ contains all possible elementary routes.

The column pool $\mathcal{P} \subseteq P$ starts with a small initial subset (e.g., single-customer routes) and grows iteratively by adding promising elementary routes from the pricing subproblem. At each iteration, the restricted master problem is solved over the current $\mathcal{P}$ to find the optimal combination of routes.

For each $p \in \mathcal{P}$, let $a_{ip} \in \{0, 1\}$ indicate whether route p serves customer $i \in C$, let c_p be the net electricity cost of p, computed as:

$$c_p = \sum_{i \in S \cup \{0\}} \sum_{t \in T} \left(p_t^+ \, P^+ \ell_{it} - p_t^- \, P^- \ell_{it} \right) \; + \; p_t^+ \, (Q - B_d)$$

The first term captures net charging/discharging costs along the route, while the second term enforces the mandatory fill-to-full charging at destination d based on arrival SOC B_d.

The original RMP reads:

$$\min \quad \sum_{p \in \mathcal{P}} c_p \, \lambda_p \tag{3.1}$$

$$\text{s.t.} \quad \sum_{p \in \mathcal{P}} a_{ip} \, \lambda_p \; = \; 1, \qquad \forall \, i \in C, \tag{3.2}$$

$$\sum_{p \in \mathcal{P}} \lambda_p \; \leq \; K_{\max}, \tag{3.3}$$

$$\lambda_p \in \{0, 1\}, \qquad \forall \, p \in \mathcal{P} \tag{3.4}$$

For column generation, we solve the linear programming relaxation of the original RMP by relaxing the integrality $\lambda_p \in \{0, 1\}$ to $\lambda_p \geq 0$. This provides a valid lower bound at each search-tree node and dual prices that drive the pricing subproblem. The relaxation master problem reads:

$$\min \quad \sum_{p \in \mathcal{P}} c_p \, \lambda_p \tag{3.5}$$

$$\text{s.t.} \quad \sum_{p \in \mathcal{P}} a_{ip} \, \lambda_p \; = \; 1, \qquad \forall \, i \in C, \tag{3.6}$$

$$\sum_{p \in \mathcal{P}} \lambda_p \; \leq \; K_{\max}, \tag{3.7}$$

$$\lambda_p \; \geq \; 0, \qquad \forall \, p \in \mathcal{P} \tag{3.8}$$

Solving (3.5)–(3.8) by simplex returns the primal solution λ and the dual variables π_i for (3.6) and $\mu \; \geq \; 0$ for (3.7). Economically, π_i is the shadow value of "cover customer i once"; it is free in sign because (3.6) is an equality. The multiplier μ measures the marginal value of one additional vehicle; it is nonnegative.

The dual value of (3.5)–(3.8) is calculated by:

$$\max \quad \sum_{i \in C} \pi_i \; + \; K_{\max} \mu \tag{3.9}$$

$$\text{s.t.} \quad \sum_{i \in C} a_{ip} \, \pi_i \; + \; \mu \; \leq \; c_p, \qquad \forall \, p \in \mathcal{P}, \tag{3.10}$$

$$\pi_i \in \mathbb{R}, \qquad \forall \, i \in C, \tag{3.11}$$

$$\mu \geq 0 \tag{3.12}$$

Pricing Subproblem

Given the dual variables (π_i, μ) from the RMP LP relaxation (3.5)–(3.8), the pricing subproblem minimizes the reduced cost $\bar{c}_p$ (3.13) of an elementary route $p \in P$ to find one or more routes with $\bar{c}_p < 0$ for addition to the column pool $\mathcal{P}$. A negative reduced cost indicates that the route improves

the RMP objective by reducing the total cost **desaulniers2005column**. The reduced cost is:

$$\bar{c}_p = c_p - \sum_{i \in C} \pi_i a_{ip} - \mu, \tag{3.13}$$

where c_p is the net electricity cost of route p as defined in the RMP.

This problem is formulated as an *Elementary Shortest Path Problem with Resource Constraints (ESPPRC)* **irnich2005shortest**, **feillet2004exact**, which seeks the shortest elementary path from origin 0 to destination d with time and energy resource constraints. The "elementary" property ensures each customer is visited at most once to maintain set-partitioning compatibility with the RMP. The ESPPRC model is:

$$\min \quad \sum_{i \in S \cup \{0\}} \sum_{t \in T} \left(p_t^+ P^+ \ell_{it} - p_t^- P^- \ell_{it} \right) + p_t^+ (Q - B_d) - \sum_{i \in C} \pi_i x_i - \mu \tag{3.14}$$

where $x_i = 1$ if customer i is visited, and the objective minimizes $\bar{c}_p$.

The pricing route must satisfy the feasibility constraints adapted from (2.2)–(2.14) for a single route, with vehicle index k removed:

- *Routing constraints*:

$$\sum_{j:(i,j) \in A} x_{ij} = 1, \qquad \forall i \in C, \tag{3.15}$$

$$\sum_{j:(0,j) \in A} x_{0j} \leq 1, \qquad \sum_{i:(i,d) \in A} x_{id} \leq 1, \tag{3.16}$$

$$\sum_{j:(i,j) \in A} x_{ij} = \sum_{h:(h,i) \in A} x_{hi}, \qquad \forall i \in V \setminus \{0, d\}, \tag{3.17}$$

- *Time constraints*:

$$T_j \geq T_i + \tau_{ij} - M_T (1 - x_{ij}), \qquad \forall (i,j) \in A, \tag{3.18}$$

$$a_i \leq T_i \leq b_i, \qquad \forall i \in C \cup \{0, d\}, \tag{3.19}$$

$$\sum_{t \in T} \ell_{it} \leq W_i, \qquad \forall i \in E, \tag{3.20}$$

$$\ell_{it} \leq \delta - \sigma_{it} + M_T (1 - z_i^+), \qquad \forall i \in E, t \in T, \tag{3.21}$$

- *Power and mode exclusivity*:

$$0 \leq \ell_{it} \leq \bar{\ell} z_i^+, \quad 0 \leq \ell_{it} \leq \bar{\ell} z_i^-, \quad z_i^+ + z_i^- \leq 1, \qquad \forall i \in E, t \in T, \tag{3.22}$$

- *Energy constraints*:

$$B_i^{\mathsf{dep}} = \begin{cases} B_i + \sum_{t\in T} E_{it}, & i \in E, \\ B_i, & i \in C, \end{cases} \qquad \forall i \in V, \tag{3.23}$$

$$B_j \geq B_i^{\mathsf{dep}} - e_{ij} - M_B(1 - x_{ij}), \qquad \forall (i,j) \in A, \tag{3.24}$$

$$0 \leq B_i, B_i^{\mathsf{dep}} \leq Q, \qquad \forall i \in V, \tag{3.25}$$

$$\sum_{j:(0,j)\in A} x_{0j} = \sum_{i:(i,d)\in A} x_{id}, \quad B_0 = Q, \quad B_d^{\mathsf{dep}} = Q, \tag{3.26}$$

- *Variable domains*:

$$x_{ij}, z_i^+, z_i^- \in \{0,1\}, \quad T_i, W_i, \sigma_{it}, \ell_{it}, E_{it}, B_i, B_i^{\mathsf{dep}} \geq 0, \tag{3.27}$$

5.2 Labeling Algorithm for the Pricing ESPPRC

State, resources, and initialization

We solve the pricing ESPPRC with a forward label-setting algorithm. Labels are extended from the depot while propagating time, state of charge, and the visited-customer set to maintain elementarity. Routes with negative reduced cost $\bar{c}_p$ per are added to the RMP to improve the objective (cite).

$$a_u \leq T \leq b_u, \quad 0 \leq B \leq Q, \quad T + \mathsf{LB}_{\mathsf{toDep}}(u) \leq b_d, \tag{3.28}$$

where $\mathsf{LB}_{\mathsf{toDep}}(u)$ is the shortest travel time from u directly to depot d over feasible arcs. We use $\mathsf{LB}_{\mathsf{toDep}}(u)$ as an admissible bound in the label-setting algorithm: a label at u is extended only if its current resources plus $\mathsf{LB}_{\mathsf{toDep}}(u)$ still permit reaching d within the time-window and energy constraints; otherwise the label is discarded, ensuring that every surviving label can be completed into a feasible route.

The potential ρ accumulates the net electricity cost minus dual benefits:

$$\rho = \sum_{i\in S\cup\{0\}} \sum_{t\in T} \left(p_t^+ P^+ \ell_{it} - p_t^- P^- \ell_{it}\right) - \sum_{i\in V} \pi_i - \mu, \tag{3.29}$$

where the vehicle-cap dual μ is subtracted at the root, as each route uses one vehicle ($\nu_p = 1$). The root label is:

$$\mathcal{L}_0 = \left(0, 0, Q, \varnothing, -\mu, \varnothing, 0\right) \tag{3.30}$$

When a label reaches destination d within $[a_d, b_d]$, we schedule charging to achieve $B_d^{\mathsf{dep}} = Q$. The required charging energy is $Q - B_d^{\mathsf{arr}}$, where B_d^{arr} is the arrival SOC at d. Let $\{\ell_{dt}\}_{t\in T}$ be the charging

durations at d satisfying $P^+ \sum_t \ell_{dt} = Q - B_d^{\text{arr}}$. The fill-to-full cost is:

$$q_d = \sum_{t \in T} p_t^+ P^+ \ell_{dt} \tag{3.31}$$

yielding the route's reduced cost $\bar{c}_p = \rho + q_d$.

For early pruning, we compute a best-case fill-to-full cost $\underline{q}_d(B_d^{\text{arr}})$ by allocating $Q - B_d^{\text{arr}}$ to the cheapest available TOU periods at d, giving the optimistic label bound:

$$\underline{\bar{c}}(\mathcal{L}) = \rho + \underline{q}_d(B_d) \tag{3.32}$$

If $\underline{\bar{c}}(\mathcal{L}) \geq 0$, the label cannot lead to a negative reduced-cost route and is discarded.

Label Extensions: To Customers and Stations

From any feasible label $\mathcal{L} = (u, T, B, V, \rho, S_{\text{vis}}, \ell)$, we extend to unserved customers or charging stations, checking time and energy feasibility and depot reachability at each step per (3.28).

For an unserved customer $j \in C \setminus V$ with arc $(u, j) \in A$, time propagates as:

$$T' = \max\{T + \tau_{uj}, a_j\} \quad T' \leq b_j \tag{3.33}$$

ensuring arrival at j respects travel time τ_{uj} and time window $[a_j, b_j]$. The max enforces waiting if arriving early. Energy propagates as:

$$B' = B - e_{uj} \quad 0 \leq B' \leq Q \tag{3.34}$$

deducting traction energy e_{uj} to ensure SOC remains feasible within $[0, Q]$. This checks if the vehicle can reach j. If $T' + \text{LB}_{\text{toDep}}(j) \leq b_d$, we create:

$$\mathcal{L}' = (j, T', B', V \cup \{j\}, \rho - \pi_j, S_{\text{vis}}, 0) \tag{3.35}$$

updating ρ with dual benefit π_j for covering j and resetting ℓ to allow station visits.

When the previous node is not a charging station ($\ell = 0$), we can extend to an unvisited charging station s that has not been visited before and is reachable from the current node u via a valid arc. This allows the vehicle to charge or discharge energy as needed. Time and energy propagate as:

$$T_s = \max\{T + \tau_{us}, a_s\} \quad T_s \leq b_s \quad B_s = B - e_{us} \quad 0 \leq B_s \leq Q \tag{3.36}$$

ensuring arrival at s respects its time window $[a_s, b_s]$ and SOC limits after consuming e_{us}. For the next hop $v \in (C \setminus V) \cup S \cup \{d\}$ (excluding $v = s$), we discretize energy decisions:

$$\mathcal{D}(s, v) = \left\{ \max\{0, e_{sv} - B_s\}, Q - B_s \right\} \tag{3.37}$$

choosing either enough energy to reach the next node or a full charge. This simplifies energy management decisions. For each $dq \in \mathcal{D}(s,v)$, dwell time and post-dwell SOC are:

$$W_s(dq) = \begin{cases} \dfrac{dq}{P^+} & dq \geq 0 \text{ (charge)} \\ \dfrac{-dq}{P^-} & dq < 0 \text{ (discharge)} \end{cases} \qquad B_s^{\text{out}} = B_s + dq \quad 0 \leq B_s^{\text{out}} \leq Q \tag{3.38}$$

determining dwell time from energy changes and updating the battery level. This ensures feasible charging or discharging. The transaction aligns with TOU periods $[a_{st}, b_{st}]$ using durations $\ell_{st} \geq 0$ (charge) and $\ell_{st}^{\text{dis}} \geq 0$ (discharge):

$$\sum_{t \in T} (\ell_{st} + \ell_{st}^{\text{dis}}) = W_s(dq) \quad T_s + \ell_{st} \in [a_{st}, b_{st}] \quad T_s + \ell_{st}^{\text{dis}} \in [a_{st}, b_{st}] \tag{3.39}$$

scheduling energy transactions within valid pricing periods to follow TOU rules. With mutual exclusivity:

$$\ell_{st} \cdot \ell_{st}^{\text{dis}} = 0 \quad \forall t \in T \tag{3.40}$$

ensuring charging and discharging do not occur simultaneously in any period. This maintains operational consistency. The station cost increment is:

$$q_s(dq) = \sum_{t \in T} \left(p_t^+ P^+ \ell_{st} - p_t^- P^- \ell_{st}^{\text{dis}} \right) \tag{3.41}$$

calculating costs based on time-of-use prices for charging or discharging. This updates the route's total cost. After dwell, propagate to v:

$$T' = \max\{T_s + W_s(dq) + \tau_{sv}, a_v\} \quad T' \leq b_v \quad B' = B_s^{\text{out}} - e_{sv} \quad 0 \leq B' \leq Q \tag{3.42}$$

verifying the next node's time window and battery level after dwell and travel. This ensures feasible continuation. If feasible and $T' + \text{LB}_{\text{toDep}}(v) \leq b_d$, we create:

$$\mathcal{L}' = (v, T', B', V, \rho + q_s(dq), S_{\text{vis}} \cup \{s\}, 1) \tag{3.43}$$

recording the station visit and preventing consecutive station stops. This tracks visited stations accurately. Candidate v can be prioritized by a "near-and-early" score for efficiency.

When a label reaches d, we compute the fill-to-full cost q_d using (3.31), yielding $\bar{c}_p = \rho + q_d$. If $\bar{c}_p < 0$, the route is returned to the RMP.

Dominance, Reachability, and Retention

To control label explosion, we apply Pareto dominance and reachability screening for labels at the same node and customer set, using a three-dimensional comparison key:

$$\Psi(\mathcal{L}) = \big(T, -B, \rho\big) \tag{3.44}$$

comparing arrival time, negative battery level, and cost potential for dominance checks. For two labels at the same node and customer set, we discard $\mathcal{L}_2$ if:

$$T_1 \leq T_2 \quad B_1 \geq B_2 \quad \rho_1 \leq \rho_2 \tag{3.45}$$

with at least one strict inequality. This ensures $\mathcal{L}_1$ can replicate any feasible extension of $\mathcal{L}_2$ with earlier or equal arrival, higher or equal battery, and lower or equal cost. We also use a depot reachability bound:

$$\eta(\mathcal{L}) = T + \mathsf{LB}_{\mathsf{toDep}}(u) \tag{3.46}$$

discarding labels unable to reach the depot within its time window. Labels with $\eta(\mathcal{L}) > b_d$ are removed. Additionally, the optimistic cost bound $\bar{c}(\mathcal{L}) = \rho + \underline{q}_d(B_d)$ from (3.32) eliminates labels that cannot produce a negative reduced cost even with the cheapest depot charging. Small numerical tolerances are applied to time and battery tests to prevent boundary issues. These rules keep the label set compact while preserving all routes with negative reduced cost, ensuring efficient pricing convergence.

5.3 Branch-and-Price: Node Algorithm and Ryan–Foster Branching

Directly solving the MILP faces two challenges. First, the route space is exponential: combining time, energy, TOU pricing, and fill-to-full at the depot into one model creates a weak LP relaxation with too many columns and constraints. Second, enforcing integer solutions is costly: the binary structure of customer coverage and route selection causes symmetry and poor bounds. A Dantzig–Wolfe (DW) reformulation splits the problem into a set-partitioning master problem over routes and a pricing subproblem, moving spatio–temporal–energy constraints to the subproblem while keeping only coverage and optional fleet size constraints in the RMP. This preserves the model's fidelity (including TOU charging and fill-to-full) while reducing dimensionality. However, the RMP LP relaxation is often fractional, so column generation alone provides only a lower bound. Branch-and-Price (B&P) tightens this bound via column generation and uses Ryan–Foster (RF) branching to enforce integer solutions. RF branches on whether a customer pair is served on the same route, adding no new RMP constraints and preserving the DW structure without altering the pricing subproblem or removing any integer-feasible solution.

We state three correctness claims for our framework.

Proposition 1 (DW equivalence and lower bound validity). Let $\mathcal{B}$ denote the branch state (sets of must-together customer groups and must-separate customer pairs). Let $P(\mathcal{B})$ be all elementary feasible routes from depot start 0 to end d satisfying $\mathcal{B}$. Then: (i) any integer-feasible solution under $\mathcal{B}$ corresponds one-to-one to a $\{0,1\}$ RMP solution over $P(\mathcal{B})$; (ii) the RMP LP value $z^{\mathsf{LP}}(\mathcal{B})$ is a valid lower bound for the optimal value under $\mathcal{B}$.

Proof sketch. (i) Each vehicle route forms an RMP column with coverage $= 1$ and, optionally, $\sum \nu_p \leq K_{\mathsf{max}}$. Any $\{0,1\}$ RMP solution reconstructs routes, with station transactions and fill-to-full at d ensured by Sections 2 and 3.2 feasibility. The branch state filters routes without affecting this mapping. (ii) The LP relaxation never exceeds any integer solution's value, ensuring a valid lower bound.

Proposition 2 (RF branching effectiveness). For any customer pair $\{i, j\}$, RF branching into Separate (different routes) and Together (same route) covers all integer possibilities without excluding any optimal solution.
Proof sketch. In any integer solution, $\{i, j\}$ are served on the same or different routes. RF branching filters routes without adding RMP constraints, leaving the pricing subproblem unchanged per Section 3.2. Thus, no integer optimum is removed.

Proposition 3 (Global convergence). Track the best integer solution $z^\star$ (upper bound) and the minimum LP lower bound LB over open nodes. When no negative reduced-cost columns remain and the gap $\text{gap} = (z^\star - \text{LB})/\max\{1, |z^\star|\}$ meets the termination criterion, $z^\star$ is globally optimal.
Proof sketch. Each node LP provides a valid lower bound. When all nodes are pruned ($z^{\text{LP}}(\mathcal{B}) \geq z^\star$) or fathomed ($z^{\text{IP}}(\mathcal{B}) \approx z^{\text{LP}}(\mathcal{B})$), the incumbent $z^\star$ is optimal.

In a branch node, we solve a column-restricted RMP under branch state $\mathcal{B}$, perform column generation until convergence, then solve a binary set-partitioning problem over the current column pool for a node upper bound. The branch state includes pairwise constraints (same-route or different-route), restricting the admissible route set:

$$P(\mathcal{B}) = \big\{p \in P : p \text{ satisfies all together/separate rules in } \mathcal{B}\big\} \tag{3.47}$$

defining routes that comply with the branch state's pairing rules. The node RMP, identical to Section 3.1 but restricted to $P(\mathcal{B})$, yields the node lower bound $z^{\text{LP}}(\mathcal{B})$. We enforce (3.47) by filtering: routes violating $\mathcal{B}$ (e.g., including a must-separate pair or missing a must-together group) are discarded. To ensure feasibility in early Together branch iterations, we add high-cost artificial columns covering required groups with cost M, used only if no physical route exists.

The node workflow starts with an RMP initialized with artificial columns and inherited columns from parent/sibling nodes. Solve the LP relaxation to get duals (π, μ), then call the pricing subproblem (Section 3.2) to add negative reduced-cost columns, iterating until none remain under $\mathcal{B}$. The resulting LP value is $z^{\text{LP}}(\mathcal{B})$. Then, set λ_p (route weight) to binary and solve the set-partitioning IP for the node upper bound $z^{\text{IP}}(\mathcal{B})$. Across the tree, we track:

$$\text{LB} = \min_{\text{open nodes}} z^{\text{LP}}(\mathcal{B}) \quad \text{UB} = z^\star \quad \text{gap} = \frac{\text{UB} - \text{LB}}{\max\{1, |\text{UB}|\}} \tag{3.48}$$

monitoring the global lower bound, upper bound, and convergence gap. A node is pruned if $z^{\text{LP}}(\mathcal{B}) \geq z^\star$ or fathomed if $z^{\text{IP}}(\mathcal{B}) - z^{\text{LP}}(\mathcal{B})$ is within tolerance. Otherwise, select a customer pair for RF branching and recurse.

The RF branching signal is the co-routing probability. Let λ_p be the LP weight of route p, and $a_{ip} \in \{0, 1\}$ indicate if route p serves customer i. For a pair $\{i, j\}$, define:

$$\sigma_{ij} = \sum_{p \in P(\mathcal{B})} \lambda_p \, \mathbb{I}\{a_{ip} = 1 \wedge a_{jp} = 1\} \tag{3.49}$$

measuring the likelihood that i and j share a route in the LP solution. Values of σ_{ij} far from 0 or 1

indicate fractional solutions. Among pairs with $\varepsilon < \sigma_{ij} < 1 - \varepsilon$, RF selects the pair minimizing $|\sigma_{ij} - \frac{1}{2}|$ and creates two branches: Separate, forbidding co-routing:

$$P(\mathcal{B}^-) = \{p \in P(\mathcal{B}) : \neg(a_{ip} = 1 \wedge a_{jp} = 1)\} \tag{3.50}$$

excluding routes serving both i and j; and Together, enforcing co-routing:

$$P(\mathcal{B}^+) = \{p \in P(\mathcal{B}) : a_{ip} = a_{jp}\} \tag{3.51}$$

requiring routes to serve i and j together or neither. Both branches use column filtering, preserving RMP sparsity. For the Together branch, an artificial column covering $\{i, j\}$ with cost M ensures early feasibility.

We use depth-first search to quickly find good incumbents and enable pruning. Each child inherits the parent's filtered column pool, reusing pricing results. RF branching only modifies $P(\mathcal{B})$, leaving pricing feasibility and costs (including TOU charging and fill-to-full) unchanged. At any node, the RMP LP provides a valid lower bound. When no open node can improve the bound and the gap meets the stopping rule, $z^\star$ is optimal. RF branching eliminates fractional overlap by enforcing same-route or different-route decisions, driving the solution toward integer optima with stable convergence.

5.4 Initialization and Minimal Controls

At each branch node, we initialize the Restricted Master Problem with artificial routes and an inherited route pool. Given the branch state $\mathcal{B}$, which specifies must-together customer groups and must-separate pairs, we create an artificial route p_G^{art} for each must-together group G to cover all customers in G. For each uncovered singleton customer i, we add an artificial route p_i^{art}. These routes have a high cost M to ensure early feasibility. The set of artificial routes is:

$$\mathcal{P}^{\text{art}}(\mathcal{B}) = \{p_G^{\text{art}} : G \in \text{together-groups}(\mathcal{B})\} \cup \{p_i^{\text{art}} : i \in C \setminus (\cup G)\} \tag{3.52}$$

defining routes to cover required customer groups or singletons.We also use the route pool inherited from parent or sibling nodes, selecting only those routes that satisfy the branch state's pairing rules through a filtering process. The initial route set is:

$$\mathcal{P}^0(\mathcal{B}) = \left(\mathcal{P}^{\text{art}}(\mathcal{B}) \cup \mathcal{P}^{\text{pool-in}}\right) \cap P(\mathcal{B}) \tag{3.53}$$

combining artificial and inherited routes that comply with $\mathcal{B}$, where $P(\mathcal{B})$ is the set of admissible elementary routes per (3.47). The RMP is solved using Gurobi's primal simplex method (Method=0) for efficient incremental route additions.

For pricing, only routes with negative reduced cost are added to the RMP. We use a small positive threshold ε_{rc} (e.g., 10^{-6}), exposed as a solver parameter `RC_TOL`, to ensure numerical stability. A

route p is added if

$$\bar{c}_p \leq -\varepsilon_{\text{rc}}. \tag{3.54}$$

Column generation (CG) stops at a node when no route satisfies (3.54).

5.5 K-best Multi-Route Retrieval and Adaptive K

Unlike standard CG, which adds one negative reduced-cost route ($K = 1$), we retrieve up to K such routes per pricing iteration. This: (i) captures diverse routes for a given dual solution, reducing RMP iterations; (ii) stabilizes dual prices faster, improving pricing efficiency; and (iii) pairs well with dominance and pruning rules. The K-best approach preserves optimality, as the stopping criterion (no negative reduced-cost routes) ensures correct lower bounds and solutions, only varying the number of routes added per iteration.

We use an adaptive K mechanism, adjusting K within $[K_{\text{min}}, K_{\text{max}}]$ based on recent route activity and pricing load. Let $t_{\text{RMP}}^{(r)}$ and $t_{\text{PR}}^{(r)}$ be the RMP and pricing computation times in iteration r. The pricing load ratio is:

$$\rho^{(r)} = \frac{t_{\text{PR}}^{(r)}}{t_{\text{RMP}}^{(r)} + t_{\text{PR}}^{(r)}} \tag{3.55}$$

measuring the proportion of time spent on pricing. Let $\{a^{(r-w)}\}_{w=0}^{W-1}$ be the number of new routes added over a window $W = \texttt{ADAPT\_WIN}$. The update rule for K is:

$$K^{(r+1)} = \begin{cases} \min\{K_{\text{max}}, K^{(r)} + 2\} & \text{if } \left(\forall w \in [0, W-1], a^{(r-w)} \leq 1\right) \\ \max\{K_{\text{min}}, K^{(r)} - 1\} & \text{if } \rho^{(r)} > \rho_{\text{high}} \text{ and } a^{(r)} \geq 1 \\ K^{(r)} & \text{otherwise} \end{cases} \tag{3.56}$$

adjusting K based on pricing load and route addition trends, with $\rho_{\text{high}} = \texttt{PRICING\_LOAD\_HIGH} = 0.70$, a threshold to detect high pricing effort. In experiments, the baseline uses standard column generation ($K = 1$), while the enhanced setting uses K-best with adaptive K. Other solver settings and tolerances remain identical.

Generated routes are added to the pool and reused in descendant nodes after $\mathcal{B}$-filtering. We exclude near-duplicate routes (same customer sets, path patterns, or end-of-route battery level up to rounding) to limit pool growth. Deterministic random seeds ensure reproducibility, and CSV loggers support post-hoc analysis without affecting decisions.

6 Computational Study

All experiments are conducted on a uniform hardware platform to ensure comparability: an AMD Ryzen 7 5800H (8 cores, 16 threads, 3.20 GHz), 16 GB RAM, and an NVIDIA GeForce RTX 3060 Laptop GPU (6 GB) running Windows 11. Unless otherwise specified, all methods adhere to a wall-clock time limit of 3600 seconds, utilize identical thread counts, and employ numerical tolerances

$\texttt{FeasTol} = \texttt{IntTol} = 10^{-6}$ and a pricing threshold $\texttt{RC\_TOL} = 10^{-6}$. Within this fixed-time framework, we monitor the best feasible objective value (upper bound) and the current lower bound throughout the computation, enabling direct comparison of solution quality at any termination point.

Benchmark EVRPTW instances are adapted to incorporate a three-segment time-of-use (TOU) pricing scheme with constant buy/sell prices within each segment. Vehicles depart the origin depot fully charged and are recharged to full capacity upon returning to the destination depot. To evaluate scalability, instances are grouped by customer size, with aggregated statistics presented in the main text; detailed group boundaries and instance mappings are provided in the data appendix. We compare three methods under consistent modeling and pruning conditions: (i) a direct MILP baseline solved by Gurobi; (ii) a branch-and-price variant without multiple column returns or adaptive control (*B&P–NoK*), which returns only the first negative reduced-cost column ($K = 1$); and (iii) a branch-and-price variant with K-best and adaptive K mechanisms (*B&P–K*), which injects multiple negative reduced-cost columns per pricing iteration and dynamically adjusts K based on recent column additions and the proportion of time spent in pricing. The *B&P–K* approach empirically accelerates convergence of the upper bound within the time limit.

6.1 Overall Performance and Per-Instance Results

We evaluate the performance of three methods under a unified wall-clock time limit and consistent numerical settings to ensure comparisons reflect algorithmic differences rather than hardware or tolerance variations. The relative gap is calculated as

$$\text{gap} = \frac{\text{UB} - \text{LB}}{\max\{1, |\text{UB}|\}} \times 100\%,$$

where UB represents the best feasible objective value obtained and LB denotes the current global lower bound. If a method fails to find a feasible solution within the time limit, we report the best available bound(s) and explicitly indicate the timeout event. Unless otherwise specified, grouped statistics are based on the same time limit and thread configuration.

We present size-grouped performance statistics for the three methods (Gurobi, *B&P–NoK*, and *B&P–K*) in Table 4. For each size group, we report the instance identifiers, followed by the mean and standard deviation (in parentheses) of the best upper bound (UB), best lower bound (LB), relative gap (%), and runtime (seconds). Median and 90th percentile values are included in the table footnote for robustness.

We provide a detailed per-instance comparison in Table 5. For the Gurobi baseline, we report the best feasible objective (UB), best lower bound (LB), relative gap (%), total wall-clock time (seconds), and a timeout flag. The branch-and-price variants (*B&P–NoK* and *B&P–K*) additionally include the number of explored nodes, column generation iterations, pricing time share (pricing time divided by total time), and total generated columns. These metrics enable analysis of whether performance differences stem from the speed of generating feasible solutions, bound tightening, or time allocation between the master problem and pricing.

Table 4. Grouped performance under a unified time limit (filled with current logs).

Size	Instance	Method	UB (mean±sd)	LB (mean±sd)	Gap % (mean±sd)	Time (s) (mean
Small	—	Gurobi	—	—	—	—
	C101C5	B&P–NoK	39.137±1.755	39.137±1.755	0.00±0.00	299.985±297.
	C101C10	B&P–K (top 5)	39.137±1.755	39.137±1.755	0.00±0.00	104.638±103.
Medium	—	Gurobi	—	—	—	—
		B&P–NoK	—	—	—	—
		B&P–K	—	—	—	—
Large	—	Gurobi	—	—	—	—
		B&P–NoK	—	—	—	—
		B&P–K	—	—	—	—

Notes: UB/LB are means across the listed instances; gap is $(\text{UB} - \text{LB})/\max\{1, |\text{UB}|\} \times 100\%$. Time is computed as LP_sum + pricing_sum from solver logs. Additional groups will be filled as more instances are run.

Table 5. Per-instance results under the same time limit (filled where logs are available).

Inst	Method	UB	LB	Gap%	Time(s)	TO	Nodes	CG iters	Pricing loa
C101C5	B&P–NoK	37.381490	37.381490	0.00	2.60	0	—	25	0.996
	B&P–K (top 5)	37.381490	37.381490	0.00	1.543	0	—	9	0.998
C101C10	B&P–NoK	40.891916	40.891916	0.00	597.37	0	—	175	0.9999
	B&P–K (top 5)	40.891916	40.891916	0.00	207.733	0	—	—	0.9998

TO = timeout flag (1 if time limit reached, else 0). Pricing load = pricing_sum / (LP_sum + pricing_sum). Where the log does not report nodes/columns, we leave "—" as a placeholder.

6.2 Solution Quality Over Time, Time Composition, and Scalability

To evaluate the performance of the three methods during the optimization process, we monitor the best feasible objective value (upper bound, UB) and the current global lower bound (LB) along a common time axis. The relative gap at time t is calculated as

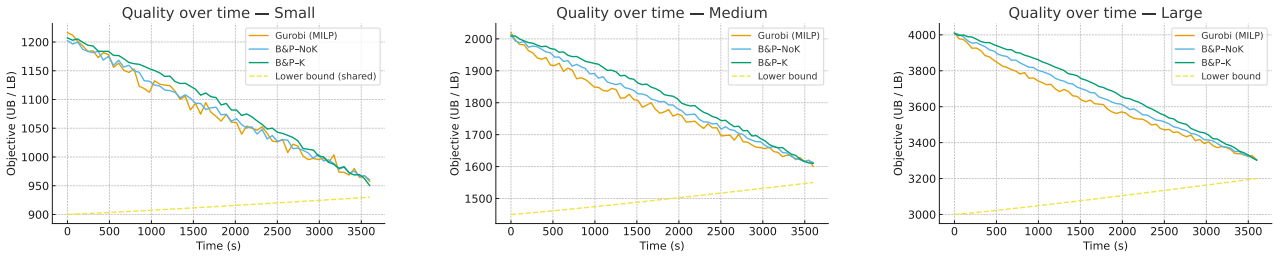
$$\text{gap}(t) = \frac{\text{UB}(t) - \text{LB}(t)}{\max\{1, |\text{UB}(t)|\}} \times 100\%.$$

Figure 2 illustrates the evolution of UB for three representative instances, one from each size group (Small, Medium, Large), with the shared LB plotted as a dashed line for reference. The *B&P–K* variant, which employs multi-column returns and an adaptive column count (K), achieves earlier and more sustained reductions in UB, particularly in the early and middle phases of the run. In contrast, the *B&P–NoK* variant, restricted to returning a single negative reduced-cost column per pricing iteration, generates high-quality feasible solutions more slowly and exhibits delayed bound tightening. The Gurobi baseline is competitive on small instances but often plateaus earlier on medium and large instances within the fixed time budget.

We decompose the wall-clock runtime into master problem (RMP/MIP) and pricing components, averaged across instances, to analyze computational efficiency. Figure 3 demonstrates that the *B&P–K* variant maintains a balanced time allocation: when pricing dominates the computational effort, the adaptive mechanism reduces K, and when recent iterations yield insufficient columns,

it increases K. This dynamic adjustment prevents excessive time spent in pricing at the expense of the master problem and ensures sufficient column generation to sustain progress, aligning with the faster UB convergence observed in Figure 2. Conversely, the *B&P–NoK* variant, limited to a single column per pricing call, requires more iterations to achieve comparable bound tightening, resulting in slower progress. The Gurobi baseline, relying solely on global MIP search without a pricing component, allocates all computational effort to a single process, often reaching the time limit on larger instances.

To assess scalability, we plot runtime against the number of customers under a unified time limit in Figure 4. As problem size increases, the Gurobi baseline exhibits rapid runtime growth, frequently hitting the time limit on medium and large instances. The *B&P–NoK* variant remains competitive only for small to medium instances, with runtime escalating sharply and often reaching the limit on larger instances due to its inefficient column generation. In contrast, the *B&P–K* variant maintains lower and more stable runtimes across a wider range of instance sizes, consistently achieving tighter upper bounds before the time limit. This suggests that *B&P–K* is better suited for practical applications with constrained computational budgets.



(a) Small instance **(b)** Medium instance **(c)** Large instance

Fig. 2. Quality over time: best upper bound (solid) and lower bound (dashed) for three representative instances (placeholders).

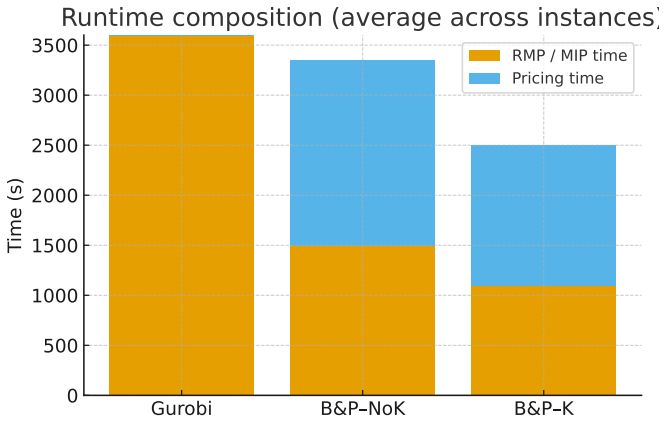


Fig. 3. Runtime composition (RMP/MIP vs. pricing), averaged across instances (placeholders).

6.3 Sensitivity Analysis and Fixed-Time Slices

We investigate the impact of the *B&P–K* configuration on runtime and solution quality, and assess the early-horizon performance of each method by measuring proximity to its final solution quality at

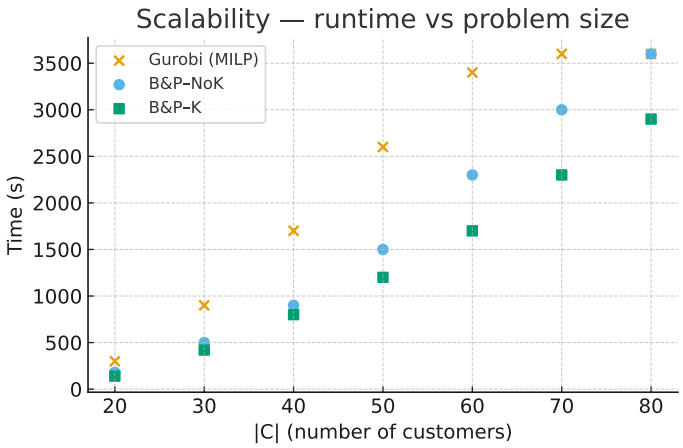


Fig. 4. Scalability: runtime vs. number of customers (placeholders).

intermediate time points. The sensitivity analysis varies the initial column return count (K_{init}), minimum return count (K_{min}), maximum return count (K_{max}), and shortlist size for successor expansion (N_{next}), while maintaining identical modeling and numerical settings across all runs.

Table 6 summarizes the results using placeholders. For each configuration, we report the wall-clock time (seconds), number of column generation iterations, pricing time share (pricing time divided by total time), and final relative gap, computed as

$$\text{gap} = \frac{\text{UB} - \text{LB}}{\max\{1, |\text{UB}|\}} \times 100\%.$$

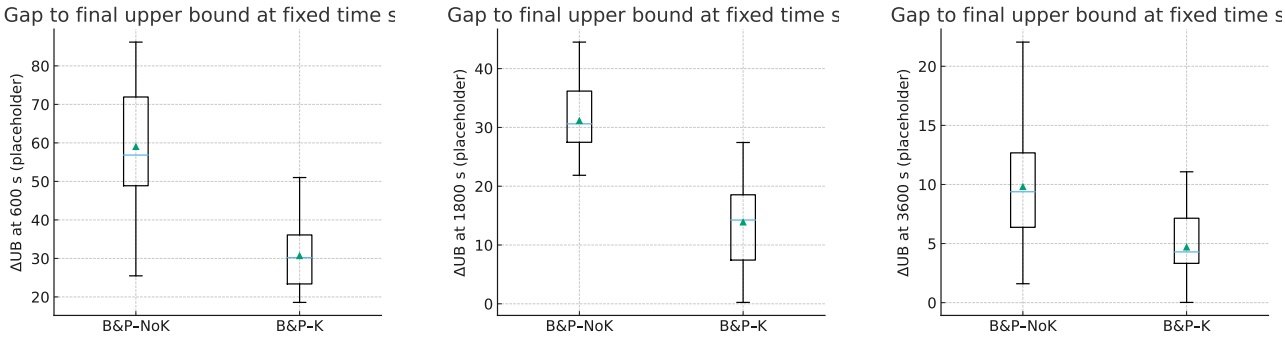
Configurations with a low return count (K) slow progress by generating insufficient columns per iteration, while excessively high counts increase pricing time disproportionately. The default adaptive configuration balances these trade-offs by adjusting K based on recent column generation activity and observed pricing time share, optimizing both runtime and solution quality.

Table 6. Sensitivity to *B&P–K* configuration (placeholders).

Setting	Time(s)	CG iters	Pricing load	Final gap%
$K_{\text{init}} = 3, K_{\text{min}} = 3, K_{\text{max}} = 8, N_{\text{next}} = 3$	—	—	—	—
$K_{\text{init}} = 5, K_{\text{min}} = 3, K_{\text{max}} = 12, N_{\text{next}} = 5$	—	—	—	—
$K_{\text{init}} = 8, K_{\text{min}} = 3, K_{\text{max}} = 16, N_{\text{next}} = 8$	—	—	—	—

Pricing load = pricing time / total time. All runs use the same time limit, thread configuration, and numerical tolerances as in Sections 5.1–5.2.

To evaluate early-horizon performance under the unified time budget, we compute the difference between the upper bound at fixed time slices $t \in \{600, 1800, 3600\}$ seconds and the final upper bound at the global time limit T, defined as $\Delta\text{UB}(t) = \text{UB}(t) - \text{UB}(T)$. Smaller $\Delta\text{UB}(t)$ values indicate earlier convergence to the final solution quality. Figure 5 presents boxplots of $\Delta\text{UB}(t)$ for the *B&P–NoK* and *B&P–K* variants, with the *B&P–K* configuration consistently showing lower medians and tighter interquartile ranges across all time slices, reflecting its superior ability to approach final solution quality early in the optimization process.



(a) 600 s **(b)** 1800 s **(c)** 3600 s

Fig. 5. Distribution of $\Delta\mathrm{UB}(t) = \mathrm{UB}(t) - \mathrm{UB}(T)$ at fixed time slices for *B&P–NoK* and *B&P–K* (placeholders).

References

[1] C. M. Martinez, X. Hu, D. Cao, E. Velenis, B. Gao, and M. Wellers, “Energy Management in Plug-in Hybrid Electric Vehicles: Recent Progress and a Connected Vehicles Perspective,” *IEEE Transactions on Vehicular Technology*, vol. 66, no. 6, pp. 4534–4549, Jun. 2017, ISSN: 1939-9359. DOI: 10.1109/TVT.2016.2582721. Accessed: Jun. 27, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7496906 (cited on p. 4).

[2] N. O. Kapustin and D. A. Grushevenko, “Long-term electric vehicles outlook and their potential impact on electric grid,” *Energy Policy*, vol. 137, p. 111 103, Feb. 2020, ISSN: 0301-4215. DOI: 10.1016/j.enpol.2019.111103. Accessed: Jun. 27, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0301421519306901 (cited on p. 4).

[3] T. Yuvaraj, K. R. Devabalaji, J. A. Kumar, S. B. Thanikanti, and N. I. Nwulu, “A Comprehensive Review and Analysis of the Allocation of Electric Vehicle Charging Stations in Distribution Networks,” *IEEE Access*, vol. 12, pp. 5404–5461, 2024, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2023.3349274. Accessed: Jun. 27, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10380227 (cited on p. 5).

[4] S. Mohanty et al., “Demand side management of electric vehicles in smart grids: A survey on strategies, challenges, modeling, and optimization,” *Energy Reports*, vol. 8, pp. 12 466–12 490, Nov. 2022, ISSN: 2352-4847. DOI: 10.1016/j.egyr.2022.09.023. Accessed: Jun. 27, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352484722017462 (cited on p. 5).

[5] S. S. Ravi and M. Aziz, “Utilization of Electric Vehicles for Vehicle-to-Grid Services: Progress and Perspectives,” en, *Energies*, vol. 15, no. 2, p. 589, Jan. 2022, Number: 2 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 1996-1073. DOI: 10.3390/en15020589. Accessed: Jun. 28, 2025. [Online]. Available: https://www.mdpi.com/1996-1073/15/2/589 (cited on p. 5).

[6] M. B. Rasheed, M. Awais, T. Alquthami, and I. Khan, “An Optimal Scheduling and Distributed Pricing Mechanism for Multi-Region Electric Vehicle Charging in Smart Grid,” *IEEE Access*, vol. 8, pp. 40 298–40 312, 2020, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2020.2976710. Accessed: Jun. 28, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9016233 (cited on p. 5).

[7] Z. Moghaddam, I. Ahmad, D. Habibi, and M. A. S. Masoum, “A Coordinated Dynamic Pricing Model for Electric Vehicle Charging Stations,” *IEEE Transactions on Transportation Electrification*, vol. 5, no. 1, pp. 226–238, Mar. 2019, ISSN: 2332-7782. DOI: 10.1109/TTE.2019.

2897087. Accessed: Jun. 28, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8632961 (cited on p. 5).

[8] F. Y. Xu and L. L. Lai, "Novel Active Time-Based Demand Response for Industrial Consumers in Smart Grid," *IEEE Transactions on Industrial Informatics*, vol. 11, no. 6, pp. 1564–1573, Dec. 2015, ISSN: 1941-0050. DOI: 10.1109/TII.2015.2446759. Accessed: Jun. 28, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7127007 (cited on p. 5).

[9] N. Bañol Arias, S. Hashemi, P. B. Andersen, C. Træholt, and R. Romero, "Distribution System Services Provided by Electric Vehicles: Recent Status, Challenges, and Future Prospects," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 12, pp. 4277–4296, Dec. 2019, Conference Name: IEEE Transactions on Intelligent Transportation Systems, ISSN: 1558-0016. DOI: 10.1109/TITS.2018.2889439. Accessed: Sep. 10, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8606438?casa_token=5LjFixAIl5cAAAAA:gOwgMwAs6nq9DIDAXtEpAkZSTQoWxKDry2I1ytY8fB2Pz2WmVzL65naB8FZ9AHJMLVD-QzqsjFQ (cited on p. 5).

[10] E. A. Grunditz and T. Thiringer, "Performance Analysis of Current BEVs Based on a Comprehensive Review of Specifications," *IEEE Transactions on Transportation Electrification*, vol. 2, no. 3, pp. 270–289, Sep. 2016, Conference Name: IEEE Transactions on Transportation Electrification, ISSN: 2332-7782. DOI: 10.1109/TTE.2016.2571783. Accessed: Sep. 10, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7476899?casa_token=Dr4hG9itIKQAAAAA:rvfP1rvSrKbzl7wJidGdA5SYrn6IiAGml8xux6UO9ZKATwZaTQ65ykugQHu0A (cited on p. 5).

[11] J. Asamer, A. Graser, B. Heilmann, and M. Ruthmair, "Sensitivity analysis for energy demand estimation of electric vehicles," *Transportation Research Part D: Transport and Environment*, vol. 46, pp. 182–199, Jul. 2016, ISSN: 1361-9209. DOI: 10.1016/j.trd.2016.03.017. Accessed: Sep. 10, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361920915300250 (cited on pp. 5, 6).

[12] D. Goeke and M. Schneider, "Routing a mixed fleet of electric and conventional vehicles," *European Journal of Operational Research*, vol. 245, no. 1, pp. 81–99, Aug. 2015, ISSN: 0377-2217. DOI: 10.1016/j.ejor.2015.01.049. Accessed: Dec. 17, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0377221715000697 (cited on p. 6).

[13] G. Lera-Romero, J. J. Miranda Bront, and F. J. Soulignac, "A branch-cut-and-price algorithm for the time-dependent electric vehicle routing problem with time windows," *European Journal of Operational Research*, vol. 312, no. 3, pp. 978–995, Feb. 2024, ISSN: 0377-2217. DOI:

10.1016/j.ejor.2023.06.037. Accessed: Sep. 11, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S037722172300509X (cited on p. 6).

[14] S. Zhang, Y. Gajpal, S. S. Appadoo, and M. M. S. Abdulkader, "Electric vehicle routing problem with recharging stations for minimizing energy consumption," *International Journal of Production Economics*, vol. 203, pp. 404–413, Sep. 2018, ISSN: 0925-5273. DOI: 10.1016/j.ijpe.2018.07.016. Accessed: Sep. 11, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925527318302810 (cited on p. 6).

[15] W. Jie, J. Yang, M. Zhang, and Y. Huang, "The two-echelon capacitated electric vehicle routing problem with battery swapping stations: Formulation and efficient methodology," *European Journal of Operational Research*, vol. 272, no. 3, pp. 879–904, Feb. 2019, ISSN: 0377-2217. DOI: 10.1016/j.ejor.2018.07.002. Accessed: Sep. 11, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0377221718306076 (cited on p. 6).

[16] R. Raeesi and K. G. Zografos, "The electric vehicle routing problem with time windows and synchronised mobile battery swapping," *Transportation Research Part B: Methodological*, vol. 140, pp. 101–129, Oct. 2020, ISSN: 0191-2615. DOI: 10.1016/j.trb.2020.06.012. Accessed: Sep. 11, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0191261520303593 (cited on p. 7).

[17] L. Tao, J. Ma, Y. Cheng, A. Noktehdan, J. Chong, and C. Lu, "A review of stochastic battery models and health management," *Renewable and Sustainable Energy Reviews*, vol. 80, pp. 716–732, Dec. 2017, ISSN: 1364-0321. DOI: 10.1016/j.rser.2017.05.127. Accessed: Sep. 11, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1364032117307736 (cited on p. 7).

[18] A. Montoya, C. Guéret, J. E. Mendoza, and J. G. Villegas, "The electric vehicle routing problem with nonlinear charging function," *Transportation Research Part B: Methodological*, Green Urban Transportation, vol. 103, pp. 87–110, Sep. 2017, ISSN: 0191-2615. DOI: 10.1016/j.trb.2017.02.004. Accessed: Sep. 11, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0191261516304556 (cited on p. 7).

[19] W. Wang and J. Zhao, "Partial linear recharging strategy for the electric fleet size and mix vehicle routing problem with time windows and recharging stations," *European Journal of Operational Research*, vol. 308, no. 2, pp. 929–948, Jul. 2023, ISSN: 0377-2217. DOI: 10.1016/j.ejor.2022.12.011. Accessed: Sep. 11, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0377221722009389 (cited on p. 7).

[20] W. K. Anuar, L. S. Lee, S. Pickl, and H.-V. Seow, “Vehicle Routing Optimisation in Humanitarian Operations: A Survey on Modelling and Optimisation Approaches,” en, *Applied Sciences*, vol. 11, no. 2, p. 667, Jan. 2021, Number: 2 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 2076-3417. DOI: 10.3390/app11020667. Accessed: Mar. 24, 2025. [Online]. Available: https://www.mdpi.com/2076-3417/11/2/667 (cited on p. 7).

[21] G. Nagy and S. Salhi, “Heuristic algorithms for single and multiple depot vehicle routing problems with pickups and deliveries,” *European Journal of Operational Research*, Logistics: From Theory to Application, vol. 162, no. 1, pp. 126–141, Apr. 2005, ISSN: 0377-2217. DOI: 10.1016/j.ejor.2002.11.003. Accessed: Mar. 24, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0377221703008361 (cited on p. 7).

[22] R. Liu, X. Xie, V. Augusto, and C. Rodriguez, “Heuristic algorithms for a vehicle routing problem with simultaneous delivery and pickup and time windows in home health care,” *European Journal of Operational Research*, vol. 230, no. 3, pp. 475–486, Nov. 2013, ISSN: 0377-2217. DOI: 10.1016/j.ejor.2013.04.044. Accessed: Mar. 24, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0377221713003585 (cited on p. 7).

[23] T. Öztaş and A. Tuş, “A hybrid metaheuristic algorithm based on iterated local search for vehicle routing problem with simultaneous pickup and delivery,” *Expert Systems with Applications*, vol. 202, p. 117 401, Sep. 2022, ISSN: 0957-4174. DOI: 10.1016/j.eswa.2022.117401. Accessed: Mar. 24, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S095741742200745X (cited on p. 8).

[24] J. Tang, C. Qi, and H. Wang, “Integrated optimization of order splitting and distribution routing for the front warehouse mode e-retailing,” *International Journal of Production Research*, vol. 0, no. 0, pp. 1–22, 2023, Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00207543.2023.2 ISSN: 0020-7543. DOI: 10.1080/00207543.2023.2200556. Accessed: Feb. 3, 2024. [Online]. Available: https://doi.org/10.1080/00207543.2023.2200556 (cited on p. 8).

[25] Y. Zang, M. Wang, and M. Qi, “A column generation tailored to electric vehicle routing problem with nonlinear battery depreciation,” *Computers & Operations Research*, vol. 137, p. 105 527, Jan. 2022, ISSN: 0305-0548. DOI: 10.1016/j.cor.2021.105527. Accessed: Mar. 24, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0305054821002653 (cited on p. 8).

[26] H. Caceres, R. Batta, and Q. He, “Special need students school bus routing: Consideration for mixed load and heterogeneous fleet,” *Socio-Economic Planning Sciences*, vol. 65, pp. 10–19, Mar. 2019, ISSN: 0038-0121. DOI: 10.1016/j.seps.2018.02.008. Accessed: Mar. 24, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0038012117301374 (cited on p. 9).

[27] C. Yao, S. Chen, M. Salazar, and Z. Yang, “Joint Routing and Charging Problem of Electric Vehicles With Incentive-Aware Customers Considering Spatio-Temporal Charging Prices,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 11, pp. 12 215–12 226, Nov. 2023, ISSN: 1558-0016. DOI: 10.1109/TITS.2023.3286952. Accessed: Jun. 28, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10164169 (cited on pp. 9, 10).

[28] J. Barco, A. Guerra, L. Muñoz, and N. Quijano, “Optimal Routing and Scheduling of Charge for Electric Vehicles: A Case Study,” en, *Mathematical Problems in Engineering*, vol. 2017, no. 1, p. 8 509 783, 2017, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1155/2017/8509783, ISSN: 1563-5147. DOI: 10.1155/2017/8509783. Accessed: Jun. 28, 2025. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1155/2017/8509783 (cited on pp. 9, 10).

[29] W. Tang, S. Bi, Y. J. Zhang, and X. Yuan, “Joint Routing and Charging Scheduling Optimizations for Smart-Grid Enabled Electric Vehicle Networks,” in *2017 IEEE 85th Vehicular Technology Conference (VTC Spring)*, Jun. 2017, pp. 1–5. DOI: 10.1109/VTCSpring.2017.8108290. Accessed: Jun. 28, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8108290 (cited on pp. 9, 10).

[30] A. Triviño-Cabrera, J. A. Aguado, and S. d. l. Torre, “Joint routing and scheduling for electric vehicles in smart grids with V2G,” *Energy*, vol. 175, pp. 113–122, May 2019, ISSN: 0360-5442. DOI: 10.1016/j.energy.2019.02.184. Accessed: Jun. 28, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360544219303901 (cited on pp. 9, 10).

[31] P. Liu et al., “Joint Route Selection and Charging Discharging Scheduling of EVs in V2G Energy Network,” *IEEE Transactions on Vehicular Technology*, vol. 69, no. 10, pp. 10 630–10 641, Oct. 2020, ISSN: 1939-9359. DOI: 10.1109/TVT.2020.3018114. Accessed: Jun. 28, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9172098 (cited on pp. 9, 10).

[32] B. Lin, B. Ghaddar, and J. Nathwani, “Electric vehicle routing with charging/discharging under time-variant electricity prices,” *Transportation Research Part C: Emerging Technologies*, vol. 130, p. 103 285, Sep. 2021, ISSN: 0968-090X. DOI: 10.1016/j.trc.2021.103285. Accessed: Jun. 28, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0968090X21002965 (cited on pp. 9, 10).

[33] M. Schneider, A. Stenger, and D. Goeke, “The Electric Vehicle-Routing Problem with Time Windows and Recharging Stations,” *Transportation Science*, vol. 48, no. 4, pp. 500–520, Nov. 2014, Publisher: INFORMS, ISSN: 0041-1655. DOI: 10.1287/trsc.2013.0490. Accessed:

Sep. 12, 2024. [Online]. Available: https://pubsonline.informs.org/doi/abs/10.1287/trsc.2013.0490 (cited on p. 10).

Intersectional Configurations of Energy Poverty: A Three-Level Modelling Framework for England

Sara Tavakoli

1. Introduction

Energy poverty is currently among the main policy issues in the European Union (EU) due to challenges linked to the energy sector transition and housing inefficiency [1]. Recent geopolitical tensions affecting global energy supply chains, have increased energy price volatility and inflation, placing additional pressure on household energy affordability and amplifying energy poverty risks worldwide [2]. While countries in the Global South primarily struggle with energy access, affordability dominates in the Global North. Even within countries, sub-national variation in energy poverty is substantial, and regional targeting depends critically on how vulnerability is modelled.

Most regional energy poverty models treat socio-economic characteristics as independent and additive. Existing studies frequently focus on selected dimensions of energy poverty, for example analysing fuel poverty risk through housing or climatic factors [3], [4] and operationalise vulnerability through two-level structures in which an index is constructed as a weighted summation of characteristics, such as multidimensional energy poverty indices and composite heating-risk indicators [5], [6]. Additive models fail to capture situations where energy poverty risk arises from the combined presence of multiple socio-economic characteristics rather than from any single characteristic alone. For example, low income alone may not produce severe energy poverty, but when combined with poor housing efficiency and reliance on expensive off-grid heating systems, vulnerability can increase substantially which is not equal to summation of their marginal effects.

This additive assumption generates three related problems: misclassification, misallocation, and intersectional masking. Under additive structures, the presence of a single socio-economic characteristic uniformly shifts the index regardless of its interaction with other characteristics. As a result, some households are incorrectly identified as vulnerable while others are overlooked, leading to potential misallocation of policy resources.

Intersectionality framework [7] argues that the effect of one socio-economic characteristic depends on the presence of others. It offers a way to conceptualise vulnerability as conditional rather than additive. Individuals are characterised by multiple socio-economic attributes that interact, and the effect of any one socio-economic characteristic on energy poverty is contingent upon others [8]. People with similar socio-economic characteristics may experience different levels of energy poverty in different climates or dwellings, not only because of independent environmental effects, but because socio-economic and health characteristics interact differently across contexts.

To address the limitations of two-level models, we propose a three-level modelling framework using intersectionality framework. The socioeconomic characteristics like such as age, sex, ethnicity, education and income, health and housing characteristics form first level of the model. Energy poverty index, low-income low energy efficiency index used in England, is the third level. Intersectionality motivates the intermediate structural layer, which is operationalised using PCA-derived configurations of socio-economic drivers. We will apply this three-level framework with intersectionality to real regional data of Westminster Parliamentary Constituencies in England.

This paper contributes to regional studies in three ways. First, it provides a conceptual contribution by framing intersectionality as a mediating mechanism in quantitative energy poverty modelling, arguing that additive models misrepresent vulnerability because individuals occupy multiple social positions simultaneously. Second, it offers a methodological contribution by proposing a three-level framework that introduces an intermediate layer between individual characteristics and outcomes, enabling the modelling of intersectional effects. Third, it delivers a regional and policy contribution by applying this framework to England, showing that although traditional two-level models may achieve stronger predictive performance, the three-level approach offers a more interpretable representation of vulnerability, allowing more precise identification of at-risk groups and supporting better targeted policies.

2. Conceptualisation

The two-level framework in energy poverty modelling links socio-economic drivers directly to an energy poverty index. Within this framework, three main approaches exist.

First, some studies analyse only one or a few characteristics, identifying relationships with energy poverty but lacking an integrated structure suitable for regional policy analysis [9], [10]. Second, many proposed models include multiple characteristics simultaneously, typically aggregating them through weighted or geometric summation [11], [12]. Although weighting methods vary, these models assume independent and fixed marginal effects, meaning the influence of one factor does not change in the presence of others. This separability prevents the modelling of intersectional vulnerability. Third, Boolean logic models classify households based on combinations of conditions [13], but they usually produce binary or categorical outcomes, limiting the ability to capture continuous variations in vulnerability.

Overall, these additive or categorical approaches tend to ignore interaction between drivers, which can lead to misclassification of vulnerability and spatial distortion when results are aggregated to regional levels. For example, households with children typically have higher energy needs due to increased heating, lighting, and appliance use. However, such households may face lower energy poverty risk when adults have stable employment and sufficient income. In contrast, households of young adults who are full-

time students may have lower energy needs but still experience higher vulnerability due to limited income. Additive models that treat these characteristics independently may therefore misclassify vulnerability.

This study uses the Capability Approach [14] to distinguish between drivers of energy poverty (EP) and its outcomes. The capability approach, introduced by Sen, conceptualises poverty as deprivation of basic capabilities rather than solely as a lack of income or resources. Within this perspective, energy poverty is treated as a realised deprivation, meaning the observable inability to meet basic energy needs, while socio-economic and structural conditions represent the drivers that create vulnerability to energy poverty. This distinction is particularly useful for policy analysis because it separates the outcome of deprivation from the conditions that produce it. Approaches that focus primarily on observed outcomes, such as the financial burden of household energy costs, may identify households experiencing hardship but provide limited insight into the structural factors that generate vulnerability. Policies targeting only observed deprivation, such as subsidies, may therefore offer temporary relief without addressing underlying causes. By identifying and modelling the drivers of vulnerability separately from the EP index, the capability-based perspective enables more effective and targeted policy interventions.

The study further applies the Intersectionality framework introduced by [7], which argues that the impact of one socio-economic characteristic depends on the presence of others. Drivers such as age, health, sex, ethnicity, education, income, household composition, housing conditions, and climate interact to shape vulnerability. Additive models assume fixed and independent effects of each driver, which contradicts intersectional logic where marginal effects are conditional on other characteristics.

3. Methodology

The methodology compares two modelling procedures. The first is a three-level model, which introduces an intermediate layer capturing relationships among independent variables (IVs). The second is a traditional two-level model, where each IV directly affects the energy poverty index without an intermediate structure. Regression-based models are used in both approaches after standard data preparation.

During data preparation, missing values and dataset identifiers are checked. Socio-economic tabulations are cleaned by removing “does not apply” categories that contain only zeros, while retaining those with meaningful values. Because variables are expressed as proportions of households, one category from each block is omitted to serve as a reference group. The dataset is then split into training (80%) and validation (20%) samples for out-of-sample evaluation.

In the three-level model, independent variables are standardised and analysed using principal component analysis (PCA) with varimax rotation. While PCA is widely used in

statistical analysis to reduce dimensionality and identify underlying patterns in correlated datasets [15], it has not previously been applied to identify intersectional configurations of drivers in energy poverty modelling. In this study, PCA is particularly important because it identifies latent configurations of socio-economic characteristics that tend to co-occur across constituencies, allowing the analysis to capture intersectional patterns of vulnerability rather than treating drivers as independent marginal effects. Components are retained based on explained variance (~75%), the eigenvalue-greater-than-one rule, and a cap of eight components to maintain interpretability. PCA extracts latent configurations of drivers, representing combinations of characteristics that tend to co-occur across constituencies and reducing multicollinearity.

The resulting principal components are interpreted through their loadings and then used to estimate the relationship with the energy poverty index using beta regression, which is suitable for bounded dependent variables. Model robustness is tested using 5-fold cross-validation.

The two-level model serves as a benchmark consistent with common approaches in the literature, where independent variables directly influence the energy poverty index without an intermediate structural layer and each driver is assumed to have a separate marginal effect. The additive specification is first estimated using ordinary least squares (OLS), which reflects the conventional modelling approach used in many empirical studies. However, because the large number of correlated socio-economic indicators can lead to unstable coefficient estimates, a ridge regression specification is also estimated, with the regularisation parameter selected through cross-validation. Using ridge provides a statistically robust benchmark by stabilising coefficients in the presence of multicollinearity. This ensures that differences between the two-level and three-level models arise from the modelling structure rather than from estimation instability.

4. Data and Empirical context

The empirical analysis focuses on England at the level of Westminster Parliamentary Constituencies (n = 533). This spatial scale allows the study to analyse sub-national variation in energy poverty while aligning the modelling framework with the level at which official statistics are reported. Socioeconomic, health and housing data are obtained from the census data 2021 which publicly available[16].

The dependent variable is the official Low Income Low Energy Efficiency (LILEE) measure of energy poverty for 2021. Under LILEE, households are classified as energy poor if they both live in a dwelling with an energy efficiency rating of Band D or below and have residual income below the poverty line after housing and energy costs. The analysis uses

the percentage of energy-poor households in each constituency as a continuous outcome variable.

The independent variables represent structural drivers of energy poverty identified in the literature and consistent with the capability-based framework. Using Census 2021 data, the model includes demographic characteristics (e.g., age, sex, ethnicity, household composition), education, labour market status, housing conditions (e.g., tenure, accommodation type, bedrooms, heating systems), health indicators, and population density from the Office for National Statistics. Variables are measured as shares of households within each constituency, allowing the model to capture spatial differences in socio-economic conditions and housing-energy infrastructure. For example, the type of central heating variable (13 categories) records the proportion of households using different heating systems, reflecting variation between urban gas-based systems and rural alternative heating sources that may influence regional energy poverty patterns.

5. Results

Applying the criteria described in Section 3.2 resulted in eight principal components, explaining about 82% of the total variance, indicating that the main structural patterns of the independent variables are captured with a limited number of dimensions.

5.1. Model performance comparison

Predictive performance is compared between the two-level model (Ridge on independent variables) and three-level models based on principal components. The two-level Ridge model achieves the highest predictive accuracy (Test R^2 = 0.867, RMSE = 1.53). However, high predictive performance alone does not guarantee meaningful policy interpretation if the model does not capture the underlying structural drivers of vulnerability. Cross-validation results show similar overall predictive performance across models, although the beta regression on PCs performs slightly worse and shows greater variability. While beta regression is theoretically suitable for a bounded dependent variable, it does not substantially improve predictive performance compared with the linear PC model.

5.2. Three-level model: structural configurations

The PCA-based three-level model identifies eight components explaining 82.36% of the variance, representing combinations of socio-economic, housing, demographic, and spatial characteristics across constituencies. Rather than analysing drivers separately, the model evaluates configurations of characteristics and their relationship with energy poverty.

Permutation importance shows that only a subset of components explains most of the variation in energy poverty, suggesting that vulnerability is structured around specific relational patterns rather than independent marginal effects. For example, some drivers

such as housing size or occupancy indicators do not appear among the largest loadings in the most explanatory components. This does not imply that these factors are unimportant. Rather, their influence is partly captured through correlated socio-economic and housing characteristics that form broader configurations within the principal components. Among them, PC2 emerges as the dominant component, representing a socio-economic gradient. It contrasts constituencies with higher education, managerial occupations, and very good health against those characterised by lower qualifications, routine occupations, and poorer health, indicating that socio-economic disadvantage is a central structural factor in fuel poverty vulnerability.

The results show that energy poverty is shaped by combinations of socio-economic, housing, and demographic characteristics rather than isolated drivers. In the three-level model, the same driver can appear in multiple components, indicating that its effect depends on the context created by other characteristics. This pattern is consistent with the intersectionality perspective, which suggests that vulnerability emerges through interacting social positions rather than through separable effects of individual characteristics [7]. The findings therefore support the view that regional energy poverty is structured through relational configurations of drivers, and the three-level framework offers a clearer representation of how these combinations shape vulnerability across constituencies.,

5.3. Two level model

The two-level model, estimated using ridge regression, directly links independent variables to the energy poverty index and achieves higher predictive performance (higher R^2 and lower RMSE). However, some coefficients produce counterintuitive signs, such as positive associations between managerial occupations or outright home ownership and energy poverty. These inconsistencies suggest that modelling drivers as independent marginal effects remains sensitive to multicollinearity and competition between correlated variables. As a result, the model may identify where energy poverty is higher but provide limited insight into why it occurs.

The comparative analysis shows that predictive accuracy alone is insufficient for understanding regional vulnerability. While the two-level model performs better statistically, the three-level framework captures structural configurations of drivers, such as socio-economic gradients (PC2), life-stage and labour-market patterns (PC5), and housing or energy infrastructure structures (PC4). Different constituencies can therefore exhibit similar levels of energy poverty due to different underlying configurations, highlighting the importance of analysing interacting drivers. This implies that policy responses may need to vary across constituencies depending on the dominant configuration of drivers. For instance, in areas where the socio-economic gradient captured by PC2 plays a larger role, policies that improve long-term socio-economic conditions such as education, occupational opportunities, and health may help reduce

vulnerability. In contrast, in constituencies where life-stage and labour-market patterns dominate, more targeted measures may be needed to protect specific groups such as younger adults who are students and have limited income or unstable employment.

6. *Discussion and conclusion*

The discussion shows that regional energy poverty cannot be fully explained through additive models that treat drivers as independent factors. Instead, vulnerability emerges from relational configurations of socio-economic, housing, and demographic characteristics. Labour-market position (e.g., occupation, economic activity status, socio-economic classification) appears as the main structural axis, while housing, demographic, and ethnic characteristics shape how vulnerability manifests across regions.

The findings also refine the commonly observed rural–urban divide in energy poverty. Rather than being purely spatial, this difference reflects structural disparities in housing and energy infrastructure, as rural areas more often rely on alternative heating systems such as oil or bottled gas. These infrastructure constraints interact with socio-economic conditions, suggesting that targeted improvements in heating systems and energy access could be effective policy responses.

The analysis further identifies localised vulnerability configurations that may be overlooked in national policy frameworks. Even if these groups represent smaller populations, their specific combinations of housing conditions, employment structures, and demographic characteristics can create heightened vulnerability.

Health also shows complex, context-dependent effects. Although widely recognised as an energy poverty driver, its influence appears mainly through interactions with other factors such as age, tenure, and labour-market status rather than as an independent driver.

Overall, the results indicate that similar levels of energy poverty can arise from different structural configurations. Therefore, identifying where energy poverty occurs is insufficient without understanding why it occurs, which requires analysing interacting drivers. From a theoretical perspective, these findings support modelling vulnerability as a relational process rather than as the separable effect of individual socio-economic characteristics. In practical and policy terms, recognising these configurations enables more place-based interventions that reflect the specific socio-economic and housing conditions of each constituency rather than assuming uniform drivers across regions [17]. This suggests that policy responses should not focus exclusively on short-term alleviation measures such as subsidies, but also consider the structural drivers of vulnerability identified in the components. For example, the dominant socio-economic gradient captured by PC2 indicates that education, occupational structure, and health conditions jointly shape long-term resilience to energy poverty. Addressing these broader

socio-economic conditions may therefore complement conventional energy poverty interventions and support more sustainable reductions in vulnerability. Future research will extend this framework by developing an intersectionality-informed prediction model that can simulate the potential effects of different energy poverty policies and schemes across regions and a fairness index. This model and fairness index allow policymakers to evaluate both the effectiveness and the distributional fairness of their future plans.

References

[1] D. Streimikiene, V. Lekavičius, T. Baležentis, G. L. Kyriakopoulos, and J. Abrhám, "Climate Change Mitigation Policies Targeting Households and Addressing Energy Poverty in European Union," *Energies 2020, Vol. 13, Page 3389*, vol. 13, no. 13, p. 3389, Jul. 2020, doi: 10.3390/EN13133389.

[2] Washington, "International Energy Agency, International Monetary Fund, and World Bank Group, Joint Statement by the Heads of the IEA, IMF, and World Bank Group on Energy Security and Economic Stability," International Monetary Fund. [Online]. Available: https://www.imf.org/en/news/articles/2026/04/13/pr26117-joint-statement-by-the-heads-iea-imf-wbg?cid=em-COM-%5B04-2026%5D-Immediate-%5BEnglish%5D

[3] D. Bienvenido-Huertas, A. Pérez-Fargallo, R. Alvarado-Amador, and C. Rubio-Bellido, "Influence of climate on the creation of multilayer perceptrons to analyse the risk of fuel poverty," *Energy Build.*, vol. 198, pp. 38–60, Sep. 2019, doi: 10.1016/J.ENBUILD.2019.05.063.

[4] A. Pérez-Fargallo, C. Rubio-Bellido, J. A. Pulido-Arcas, and F. Javier Guevara-García, "Fuel Poverty Potential Risk Index in the context of climate change in Chile," *Energy Policy*, vol. 113, pp. 157–170, Feb. 2018, doi: 10.1016/J.ENPOL.2017.10.054.

[5] J. A. Kelly, J. P. Clinch, L. Kelleher, and S. Shahab, "Enabling a just transition: A composite indicator for assessing home-heating energy-poverty risk and the impact of environmental policy measures," *Energy Policy*, vol. 146, p. 111791, Nov. 2020, doi: 10.1016/J.ENPOL.2020.111791.

[6] O. S. Santillán, K. G. Cedano, and M. Martínez, "Analysis of Energy Poverty in 7 Latin American Countries Using Multidimensional Energy Poverty Index," *Energies 2020, Vol. 13, Page 1608*, vol. 13, no. 7, p. 1608, Apr. 2020, doi: 10.3390/EN13071608.

[7] K. W. Crenshaw, "Mapping the margins: Intersectionality, identity politics, and violence against women of color," *Public Nat. Priv. Violence Women Discov. Abus.*, pp. 93–118, Feb. 2013, doi: 10.2307/1229039.

[8] N. Simcock, K. E. H. Jenkins, M. Lacey-Barnacle, M. Martiskainen, G. Mattioli, and D. Hopkins, "Identifying double energy vulnerability: A systematic and narrative review of groups at-risk of energy and transport poverty in the global north," *Energy Res. Soc. Sci.*, vol. 82, p. 102351, Dec. 2021, doi: 10.1016/J.ERSS.2021.102351.

[9] M. Reuter, M. K. Patel, W. Eichhammer, B. Lapillonne, and K. Pollier, "A comprehensive indicator set for measuring multiple benefits of energy efficiency," *Energy Policy*, vol. 139, 2020, doi: 10.1016/j.enpol.2020.111284.

[10] L. Papada and D. Kaliampakos, "A Stochastic Model for energy poverty analysis," *Energy Policy*, vol. 116, pp. 153–164, 2018, doi: 10.1016/j.enpol.2018.02.004.

[11] N. Bonatz, R. Guo, W. H. Wu, and L. J. Liu, "A comparative study of the interlinkages between energy poverty and low carbon development in China and Germany by developing an energy poverty index," *ENERGY Build.*, vol. 183, pp. 817–831, 2019, doi: 10.1016/j.enbuild.2018.09.042.

[12] J. Sokołowski, P. Lewandowski, A. Kiełczewska, and S. Bouzarovski, "A multidimensional index to measure energy poverty: the Polish case," *Energy Sources, Part B Econ. Planning, Policy*, vol. 15, no. 2, pp. 92–112, Feb. 2020, doi: 10.1080/15567249.2020.1742817.

[13] E. Spiliotis, A. Arsenopoulos, E. Kanellou, J. Psarras, and P. Kontogiorgos, "A multi-sourced data based framework for assisting utilities identify energy poor households: a case-study in Greece," *ENERGY SOURCES PART B-ECONOMICS Plan. POLICY*, vol. 15, no. 2, pp. 49–71, 2020, doi: 10.1080/15567249.2020.1739783.

[14] A. Sen, "Development as Freedom (1999)," in *The globalization and development reader: Perspectives on development and global change*, 2014, ch. 525.

[15] I. T. Jollife and J. Cadima, “Principal component analysis: A review and recent developments,” *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.*, vol. 374, no. 2065, Apr. 2016, doi: 10.1098/RSTA.2015.0202/115142.
[16] Office for National Statistics, “Census 2021 data,” Office for National Statistics. [Online]. Available: https://www.ons.gov.uk/census/maps/choropleth
[17] R. Jones and S. Moisio, “Regions and the search for spatial justice: a question of capacity?,” *Reg. Stud.*, vol. 59, no. 1, Dec. 2025, doi: 10.1080/00343404.2024.2390505.

Investment decisions analysis with prospect theory: evidence from earnings conference calls

Yunze Xie

Abstract

As an important medium for providing information to investors, earnings conference calls play an important role in the stock market. Although many scholars have been dedicated to exploring the textual characteristics of conference call, the process and analysis of textual information is insufficient. The textual characteristics are often directly used to analyze the effect on the stock market, but the real feelings of investors towards them are ignored. Especially for the stock market, it is difficult for investors to act completely rationally, as their reactions to information are influenced by individual psychological characteristics. Therefore, I will introduce prospect theory to analyze the textual characteristics of earnings calls and combine these characteristics based on evidential reasoning model, digging out the information and exploring the real feelings of investors to the greatest degree. Therefore, this research will achieve accurate evaluation of calls by using prospect theory and predict stock returns combined with company financial performance based on interpretable artificial intelligence, and provide support for investor to make decisions on investment.

Keywords: earnings conference calls, textual characteristics, stock market, prospect theory, evidential reasoning, interpretable artificial intelligence

1. Literature review

Earnings conference calls have been established as an informative disclosure medium that provides incremental value-relevant information reflected in stock prices (Frankel, 1997; Bushee, 2003; Brown, 2004). Bowen (2002) pointed out that conference calls can increase the total amount of information available on companies. Companies hold earnings conference calls prior to the release of their annual or quarterly reports to show investors past-earnings performance and explain any differences with analyst forecasts. Brochet (2018) also further emphasized the role of conference calls in information transfer. The stock market is a highly information-intensive market. Because stock prices are very sensitive to information, earnings conference calls enable earnings information to be relayed to the market in advance, preventing sharp fluctuations in stock prices. In the United States (US), in particular, where financial markets are becoming increasingly sophisticated, earnings calls have a significant impact on the stock market. The stock market is full of a wide variety of information. In addition to numerical information, there is a large amount of unstructured textual information in company disclosures.

Clear and accurate understandings of textual information are significant for investors in corporate disclosures. The extraction of textual characteristics is therefore becoming increasingly important. By examining a sample of 10-Ks from 1994 to 2008, Loughran (2011) developed an alternative list of negative words, and five lists of positive, uncertainty, litigious, strong modal, and weak modal words that more accurately reflect the sentiment in financial texts. However, Loughran did not do in-depth research regarding the relationship between the tone of financial texts and stock returns. Huang (2014) took 363,952 analyst reports from the US between 1996 and 2008 as a sample, applied a plain Bayesian machine learning approach to analyse the textual information in analyst research reports and found that the textual information contained more information about a company's expected earnings than numerical information such as earnings forecasts and target prices. Textual information was receiving more and more attention and many features of textual information were

being discovered, so the effects of textual characteristics in company disclosure on the stock market cannot be ignored. Garcia (2023) applied machine learning to further study the reactions of stock prices to textual information, providing new dictionaries for positive and negative words in the financial context.

An earnings conference call consists of two parts, the first potion is a presentation by managers, which explains the company's performance for this quarter. The second portion is the Q&A session between analysts and managers, which can be divided into two part to reflect analysts' and managers' textual information separately. Matsumoto (2011) found that the Q&A phase was relatively more informative than the presentation phase. Blau (2015) demonstrated that managers do show a more positive tone during earnings calls, especially during the presentation portion when they can prepare in advance. Although it does not have a disproportionate impact on stock prices after market adjustment, it is still worthwhile to devote sufficient attention to this situation. Besides, in Q&A portion the words of analysts and managers also attracts different degrees of attention. Especially investors will pay extra attention to the attitude of analysts. Davis (2015) found that manager-specific optimism had a significant impact on the tone used in conference calls. But in the preliminary analysis of tone and daily returns, Davis do not separate analysts from managers. Moreover, Lee (2016) further pointed out that if managers followed predetermined scripts when answering questions during earnings calls to avoid disclosing bad news, analysts would later downgrade their earnings forecasts and bid-ask spreads would increase. Besides, Huang (2017) proves the role of analysts in information discovery and interpretation in conference calls, especially when managers hide information. Andrew (2021) conducted an in-depth analysis of the textual characteristics of buy-side analysts and found that the tone of the buy-side, especially for hedge fund analysts, was positively correlated with subsequent stock returns by using 81,652 conference call transcripts from 3,446 companies from 2007 to 2016. Therefore, this study will divide the conference calls into three parts to discuss separately.

In a conference call, the tone that contains both positive and negative information is the most important text characteristic. Price (2012) examined the incremental amount of information and the corresponding market reaction to the quarterly earnings call, and demonstrated that the tone of earnings calls was a significant predictor of stock prices and trading volume and that there was a significant positive relationship between positive tones and stock returns. By using Real Estate Investment Trusts (REIT) as an example, Doran (2012) examined the relationship between the tone of conference call and the contemporaneous stock price reaction, and verified that tone had significant explanatory power for abnormal returns. Besides, Jason (2018) provided a comprehensive analysis of the tone in earnings conference calls based on previous research, and discuss the relationship between tone and the intraday stock prices. Fu (2019) further examined the relationship between the tone of earnings calls and future stock price crashes, revealing to some extent the long-term informational role of conference call tone. These studies extended the empirical disclosure literature by examining unique aspects of quarterly earnings conference calls and the subsequent market reaction, and provide much inspiration for future research.

In addition to tone, researchers have also discovered many new textual characteristics. Jancenelle (2019) contends that warm-glow rhetoric can mitigate investors' negative reactions to earnings surprises, and positively moderates the relationship between earnings surprises and financial performance. Suslava (2021) studied the impact of euphemisms on investor reactions, and found that these euphemisms used by managers have a negative relationship with future abnormal returns. Call (2023) proposed humor as a very interesting textual characteristic and examined its role in conference calls. Although these textual characteristics have been proven to have effects on stock returns, there are still doubts about whether the information they contain is sufficient to support research in predicting stock prices. Only textual characteristics which contain the most textual information can accurately capture specific fluctuations in stock prices, rather than just analyzing the effects.

As the most deeply and directly perceived textual characteristic for investors, the tone which reflects the positive and negative attitudes of conference calls has always attracted the attention of scholars. By using computer aided content analysis, Price (2012) proved that conference call linguistic tone is a significant predictor of abnormal returns and trading volume. From the perspective of tone, Bochkay (2020) verified that conference calls extreme words can increases trading volume and has an impact on stock prices, especially for companies with weaker information environments. Fu (2021) pointed out the tone of earnings conference calls can predict future stock price crash risk, and revealed the long-term informational role of conference calls tone. The predictive ability of conference calls tone for stock returns has been extensively proven, however, previous literature only directly used tone to estimate stock prices, which makes it difficult to make full use of all the textual information. Therefore, I will use prospect theory to discuss investors' more authentic feelings towards earnings calls tone from the perspective of their individual psychological characteristics, and help investor to make decisions of investment.

Decision science has been a very important research topic in the field of business and management. Tversky (1979) put forward the prospect theory, which explained the individual decision making behavior from the perspective of psychology. By introducing the psychological characteristics of decision makers into the decision making process, prospect theory has received the attention of many scholars as soon as it appeared. Prior to this, expected utility theory dominated in decision science (Neumann,1944). Prospect theory refines expected utility theory from a psychological perspective, and it is more relevant to realistic decision making behavior by incorporating individual value perception factors into the model. After that, a lot of research on prospect theory has appeared, and it has grown considerably. At the beginning, it was revised from a statistical perspective. Tversky (1992) proposed the cumulative prospect theory to avoid the contradiction with the predominance of first-order stochastic in original prospect theory. Nilsson (2011) proposed a hierarchical bayesian approach to estimate the parameters of the cumulative prospect theory, and

Glöckner (2012) focused on the adjustment of parameters. Cumulative prospect theory was then widely used, up to now it is still a very important analytical method in decision-making problems.

Prospect theory is based on psychology and fully considers the individual's value perception factors. In practice, it is not possible for decision makers to have access to all decision relevant information as described by the expected utility theory, and therefore, decision making behavior cannot be viewed as fully rational. Especially in the stock market information is very complex, and investors find it difficult to grasp and utilize all the information. It is also difficult to achieve complete rationality in the processing of earnings calls information. Steele (2010) demonstrated what the minimum requirements are for determining rational choice. Campitelli (2010) developed the concept of limited rationality and emphasized the importance of the decision makers' expertise in decision making process. And Juechems (2021) further revealed the reasons for the "irrational" behavior of decision makers. The psychological characteristics of decision makers have a very important influence on the interpretation and prediction for decision making behavior. Therefore, the introduction of prospect theory can significantly improve the accuracy of research in analyzing stock prices that reflect investor behavior. Kirshner (2019) modeled optimism and overconfidence through probability weighting functions. Ciccarone (2020) discussed the relationship between market sentiment and fluctuations in economic activity through prospect theory.

The value function and weight function are very important components of prospect theory, which reflect the individuals' psychological characteristics in the form of mathematical expressions. In this study, they will be used to analyze the positive and negative attitudes of participants in conference calls. From a statistical point of view, the form of the value function has been tested many times, The power, logarithmic, negative exponential and quadratic forms of the value function had all been discussed. Among them, the form of power function has been widely used in practical problems

(De Giorgi, Hens and Rieger, 2010; Kirby, 2011; Gazioglu and Caliskan, 2011) In addition to the statistical perspective, the value function had also been improved from the view of emotion. Individuals' cognitive process can be divided into rational and emotional components. (Bracha and Brown, 2012; Mukherjee, 2011; Garcés and Finkel, 2019) Specially, this phenomenon is reflected mathematically in the dual systems model which divide value into affective and deliberative systems to calculate. (Mukherjee, 2010; Sahlin, Wallin and Persson, 2010) The dual systems model provides a new view for the development of decision theory by combining behavioral economics and neuroeconomics. (Grayot, 2020)

The weight function is highly subjective because it is based on the decision makers' real feelings about objective probabilities. (Krawczyk, 2015) Therefore, they perform differently in different field. (Bracha, 2020) The development of weight functions mainly focuses on two aspects: parameter-free approach and parameter approach. Parameter-free approaches are aim to describe the statistical characteristics of weight functions through social experiments. Kilka and Webe (2001) proposed a two-stage approach and first use parameter-free approach to explored curve shapes of weight functions. After that, more parameter-free approaches were discussed and applied. (van de Kuilen and Wakker, 2011; Chai and Ngai, 2020) By learning more and more properties of weight functions, based on parameter approach, many specific functional forms were proposed which can be applied to specific scenarios to solve practical problems. The first parameter approach was proposed by Karmarkar (1978), although this model violated many properties of weight functions from the current perspective, it defined the most widely used function form. After that, Tversky (1992) and Prelec (1998) proposed the two most important parameter approaches. On this basis, Wu and Gonzalez discussed the parameters which affect Curvature in the model. And Gonzalez and Wu (1999) further extended them to two-parameter models and analyzed the significance of each parameter. Gradually, more new techniques such as machine learning, and artificial intelligence were used in the analysis for decision science. Cavagnaro (2013) use adaptive design optimization to test several models of

weight function. Cabrera-Paniagua (2015) proposed an autonomous emotion decision making system to support the decision making process in the stock markets. And Mello (2021) also described a methodology for predicting the outcome of individuals' decision making process based on psychological and emotional perspective by using artificial intelligence techniques. In this study, I will use interpretable artificial intelligence to discuss investors' decision-making behaviour, in order to evaluate calls more accurately to support the investment decisions of investors.

2. Introduction

As an important medium for providing information to investors, earnings conference calls play an important role in the stock market. Although many scholars have been dedicated to exploring the textual characteristics of conference call, the processing and analysis of textual information is insufficient. The textual characteristics are often directly used to analyze the impact on the stock market, but the real feelings of investors towards them are ignored. Especially for the stock market, it is difficult for investors to act completely rationally, as their reactions to information are influenced by individual psychological characteristics. Therefore, I will introduce prospect theory and analyze the textual characteristics of earnings calls through value functions and weight functions, digging out the information to the greatest degree, and exploring the real feelings of investors. Therefore, this research will achieve accurate evaluation of calls using prospect theory and predict stock returns combined with company financial performance, and provide support for investor to make decisions on investment.

As the most deeply and directly perceived textual characteristics of conference calls for investors, the tone of participants contains the most textual information. In this study, based on the financial dictionaries constructed by Garcia (2023) and Loughran (2011), I will use four positive words dictionaries and four negative words dictionaries to capture all the attitudes of managers and analysts during conference calls. In addition, considering that investors pay different attentions to presentation

portion and Q&A portion, and they focus on analysts' words, I will divide the calls into three parts to discuss separately which will be integrated to obtain a comprehensive evaluation for the conference calls based on evidential reasoning.

The overarching goal of my research is to investigate investor decision-making behaviour through textual analysis and to develop a stock price prediction model that strikes a balance between interpretability and predictive accuracy. First, I construct a fully interpretable artificial intelligence model grounded in a three-level evidential reasoning framework (three-level MAKER model). This model integrates multiple sources of information, including fine-grained sentiment extracted from earnings conference calls, financial indicators, and behavioural mechanisms derived from prospect theory. The design is centred around interpretability, with each layer corresponding to a distinct stage of investor information processing.

At the first-level MAKER, I extract sentiment signals from five key textual components: the presentation, question, and answer sections of the earnings conference call, as well as the corresponding 10-K and 10-Q filings. Sentiment is captured using four complementary financial dictionaries with separate lexicons for positive and negative words. Each dictionary is treated as an independent piece of evidence, and the evidential reasoning (ER) approach is employed to integrate them, producing a sentiment probability distribution (positive, negative, uncertain) for each of the five textual segments.

At the second-level MAKER, I aggregate the sentiment distributions from the three conference call segments—presentation, question, and answer—to construct the overall sentiment profile of the earnings call. This step accounts for structural differences and possible variations in investor focus or credibility across different sections of the call. Again, evidential reasoning is used to synthesize the evidence, generating a unified sentiment probability distribution for the full conference call, which explicitly retains a probability mass for uncertainty, acknowledging the

presence of ambiguous or conflicting sentiment cues.

At the third-level MAKER, I introduce prospect theory to translate sentiment into investor evaluation. Specifically, I combine the 10-K and 10-Q documents to form a single sentimental reference point, and compare the sentiment of the earnings call against this baseline. The sentimental deviation is passed through a value function and a subjective probability weighting function to capture the evaluation of the calls. Finally, this evaluation score is integrated with structured firm-level financial variables using evidential reasoning to produce a probability forecast for stock price movements. This fusion of textual and financial information yields a final prediction that is both behaviourally informed and fully interpretable.

Overall, this fully interpretable model serves as the behavioural foundation of my research. It simulates how investors perceive, distort, and integrate qualitative financial disclosures in light of structural, linguistic, and cognitive biases. The outputs of this model are designed to reflect investor sentiment distributions, which are later used for decision-support and prediction in subsequent stages.

Second, I introduce prospect theory as a cognitive lens to detect and interpret sentiment in textual disclosures. This stage tests whether prospect theory can be effectively applied to model investor reactions to conference call narratives. The objective is to assess whether this behavioural structure enhances the predictive validity of textual sentiment in explaining stock returns and guiding investor decisions. To further deepen the behavioural foundation, I explore the use of artificial intelligence techniques to simulate and estimate the subjective probability weighting function that lies at the core of prospect theory. Instead of assuming a fixed function form, I allow the shape of weight functions to emerge from the data, using monotonic or S-shaped neural networks to approximate how investors actually perceive and distort probabilities under uncertainty. This approach enables me to empirically test whether real-world investor behaviour aligns with classical prospect theory or exhibits

deviations, such as context-specific distortions or asymmetries in probability sensitivity.

In addition, I investigate how investor sentiment toward uncertain or ambiguous information is integrated into decision-making. Specifically, I model how the unclassified or neutral content in conference calls is psychologically reallocated between positive and negative perceptions, reflecting differing attitudes toward ambiguity. By introducing these mechanisms into the valuation and decision layers of the model, I aim to capture the complex ways in which uncertainty interacts with sentiment and shapes expectations about future returns.

Finally, I extend the scope of analysis by comparing three modelling paradigms for conference call interpretation: (i) the fully interpretable evidential reasoning framework, (ii) a semi-interpretable deep learning model using BERT combined with SHAP model, and (iii) a black-box large language model represented by ChatGPT. To enhance decision robustness, I employ evidential reasoning to fuse the outputs of these three models, and further integrate this ensemble with machine learning techniques for stock price prediction. This hybrid approach is designed to combine the respective strengths of interpretability, language comprehension, and empirical accuracy.

Across all stages, the central emphasis is placed on balancing interpretability and performance—not only in sentiment extraction, but also in the downstream task of financial prediction. The study contributes to a deeper understanding of how explainable AI frameworks can be used to support investor decision-making in the presence of behavioural biases and textual uncertainty.

3. Research methodologies

3.1 Data

I obtain the full sample of U.S. public company quarterly earnings call transcripts

from Thomson Reuters via the WRDS database, covering the period from 2004 to 2024. Given the structural and regulatory particularities of financial firms, I exclude companies in the financial sector by applying standard SIC industry codes. This filtering ensures that the analysis focuses on general corporate disclosures, thus enhancing the generalizability of the results. After downloading and cleaning the raw transcripts, I merge the textual data with financial statement variables from COMPUSTAT and capital market outcomes from CRSP. In addition, to construct investor sentiment reference points based on prospect theory, I retrieve the full texts of each firm's corresponding 10-K and 10-Q filings and match them with the earnings calls by firm identifier and filing date proximity. This allows me to compare the sentiments of the conference calls with the associated financial report within the same fiscal quarter, enabling the measurement of reference-dependent sentiment deviations. After all filtering and matching procedures, the final dataset contains 62,257 earnings calls, with only one valid sample available in 2004. Due to limited availability in the early years, the sample size before 2010 is relatively small compared to more recent years.

For each transcript, I develop a comprehensive Python-based processing system to extract a range of textual and structural characteristics. The system identifies and separates the presentation, question, and answer sections of each call, attributes speaking turns to either management or analysts, and measures the sentiments of each segment using four groups financial dictionaries. I also extract features at the transcript related variables, speaker related variables and file related variables, thus constructing a rich and interpretable textual dataset.

Based on the file related variables and speaker related variables extracted from the transcripts, I divide each earnings conference call into three distinct structural components. The first component is the presentation portion, where company management—typically the CEO or CFO—reviews the firm’s recent financial performance and offers explanations for discrepancies between actual results and

market expectations. This portion often reflects the prepared narrative that the firm intends to communicate to the market. The second component is the question portion, which consists of inquiries raised by financial analysts. These questions are often critical, targeted, and focused on specific performance metrics, strategic uncertainties, or forward-looking guidance. Analysts play a gatekeeping role, and their questions often reflect the concerns of institutional investors, thereby serving as a channel for investor sentiment to surface in real time. The third component is the answer portion, in which managers respond to the analysts' questions. Unlike the presentation section, these answers are typically spontaneous and unstructured, often revealing managers' attitudes, confidence levels, and ability to handle scrutiny. As such, this portion offers rich information about management credibility and transparency.

Given that these three components differ significantly in terms of information content, communicative intention, and perceived credibility, they likely receive differential levels of attention from investors and exert heterogeneous effects on stock market reactions. Therefore, I analyze them separately by assigning component-specific weights and reliabilities in the sentiment fusion process, allowing the model to reflect the asymmetric influence of different sections on investor decision-making. For every part, I can extract the transcripts related variables. By using textual analysis I use the Garcia (2023) and Loughran (2011) financial text dictionary to count eight types words which are the number of LM&ML positive words, LM&ML negative words, LM positive words, LM negative words, ML positive words, ML negative words, ML positive binary words, and ML negative binary words. Then I can get the total number of all sentimental words, and calculate the percentage of each type of words.

To improve the accuracy and interpretability of sentiment measurement in conference calls, I introduce the concept of sentiment coverage as a correction factor that reflects the emotional richness or expressiveness of a text. Sentiment coverage is defined as the proportion of sentiment-bearing words among all words in a given text segment, and is used to adjust the final sentiment probability outputs. Formally, for each

dictionary and each text segment, I first calculate the internal proportion of each sentiment type. Specifically, I focus only on the words that carry sentimental information and exclude non-sentimental or irrelevant words. Let x_i denote the number of words in sentiment category i, and $\sum_j x_j$ be the total number of sentiment-bearing words. The internal sentiment proportion is given by:

$$P_i^{internal} = \frac{x_i}{\sum_j x_j} \tag{1}$$

This step captures the internal structure of the sentiment distribution, such that highly skewed or polarized content can be highlighted. Next, I compute the raw sentiment coverage rate as:

$$Coverage_{raw} = \frac{\sum_j x_j}{Total\ Words} \tag{2}$$

This value indicates the degree to which the text expresses sentiment overall. If the coverage is low (e.g., less than 3–5%), then the emotional signal may be weak or unreliable. To avoid overstating the sentiment of emotionally sparse text, I introduce a nonlinear activation function to compress the sentiment signal when coverage is low. Specifically, I define the coverage-based adjustment factor as:

$$\rho = \frac{1}{1+e^{-\beta(Coverage_{raw}-\theta)}} \tag{3}$$

where β controls the sensitivity (i.e., how steep the curve is), and θ is the activation threshold below which sentiment scores are significantly dampened. Following empirical calibration, I use $\beta = 10$ and $\theta = 0.05$ as recommended parameters. These values ensure that when sentiment words make up less than 5% of the total text, the sentiment signal is proportionally suppressed, aligning with the intuition that such texts may not carry meaningful emotional cues. The final sentiment probability for each category is then calculated as:

$$P_i^{final} = \rho \cdot P_i^{internal} \tag{4}$$

This final output reflects both the internal emotional structure of the text and the degree of sentiment coverage. When coverage is high, $\rho \approx 1$, and the internal distribution is fully preserved. When coverage is low, ρ\rhoρ approaches zero, resulting in conservative sentiment estimates. In practice, this process is repeated

independently for each of the four sentiment dictionary pairs (LM, ML, LM-ML, MB), each of which contains one positive and one negative word list. For each dictionary, I calculate the positive and negative probabilities using the above method. The residual probability mass is assigned to the uncertain category, representing either neutral tone or weak sentiment expression. By treating each dictionary as an independent source of evidence and calculating their sentiment probabilities in a consistent and interpretable way, I obtain four sets of sentiment distributions that can later be fused using evidential reasoning to obtain the final sentiment score. This method ensures that sentiment classification not only respects internal emotional structure, but also accounts for the overall expressiveness of the text. It prevents overinterpretation of sparse signals and provides a solid behavioural foundation for downstream modelling.

Besides, some information for firm fundamentals and capital market can be obtained from COMPUSTAT and CRSP. Given that company fundamentals data are generally published one year late, all company variables use the previous year's data in order to prevent the introduction of future information. And cumulative n-day abnormal returns start from the current earnings conference call date, where abnormal returns are calculated as the raw return minus the buy-and-hold return on the S&P 500 value-weighted market index (Huang, 2014). As a result, for firm i in quarter t, variables about firm fundamentals can be obtained,

$$LEVERAGE_{i,t} = \frac{(DLTT_{i,t}+DLC_{i,t})}{AT_{i,t}} \tag{5}$$

$$ROA_{i,t} = \frac{NI_{i,t}}{AT_{i,t}} \tag{6}$$

$$EP_{i,t} = \frac{EPSPI_{i,t}}{PRCC\_C_{i,t}} \tag{7}$$

$$BM_{i,t} = \frac{BE_{i,t}}{PRCC\_C_{i,t} \times CSHO_{i,t}} \tag{8}$$

$$SIZE_{i,t} = \log(AT_{i,t}) \tag{9}$$

For firm i in quarter t, cumulative n-day abnormal returns can be obtained,

$$BHRET\_0\_1_{i,t} = abnormal\ return\_0\_1_{i,t} \tag{10}$$

$$abnormal\ return\_0\_n_{i,t} = return\_0\_n_{i,t} - vwreted\_0\_n_{i,t} \tag{11}$$

$$BHRET\_0\_n_{i,t} = \left(1 + BHRET\_0\_(n-1)_{i,t}\right) \times \left(1 + abnormal\ return\_0\_n_{i,t}\right) - 1 \quad (12)$$

Where, $BHRET\_0\_n_{i,t}$ is the final predicted target. Considering that changes in a company's stock price may be due to some external causes, I use abnormal stock returns to analyze more clearly how stock prices are affected by earnings calls, and the cumulative abnormal returns can reflect the long-term impact of earnings calls. $LEVERAGE$, ROA, EP, BM and $SIZE$ are financial information which will be combined in the third-level MAKER.

3.2 Evidential reasoning

The evidential reasoning (ER) framework provides a mathematically grounded and interpretable approach for aggregating uncertain information from multiple sources. Originally developed by Yang and Xu (1999) and later extended into the more expressive MAKER model (Yang & Xu, 2025), this framework is well-suited for tasks combining complex information. In this study, I adopt the latest MAKER structure to integrate sentiment signals extracted from different dictionaries and document segments, enabling a nuanced fusion of linguistic and contextual cues. The key idea behind the ER approach is to treat each source of information as an "evidence body" that expresses its support for competing hypotheses. In my application, these hypotheses correspond to positive (H1), negative (H2), or unknown (H1&H2) sentiment orientations in the third-level MAKER. Each piece of evidence is characterized by a probability vector $p = (p_{H1}, p_{H2}, p_{unk})$, a weight vector $w = (w_{H1}, w_{H2}, w_{unk})$, and a reliability $r \in [0, 1]$, which captures the trustworthiness of the evidence source. Before fusion, each evidence is transformed into a discounted belief distribution through the following formula:

$$w = \frac{1}{p_{H1} \cdot w_{H1} + p_{H2} \cdot w_{H2} + p_{unk} \cdot w_{unk} + (1-r)} \quad (13)$$

$$m(H1) = w \cdot p_{H1} \cdot w_{H1} \quad (14)$$

$$m(H2) = w \cdot p_{H2} \cdot w_{H2} \quad (15)$$

$$m(\text{H1\&H2}) = w \cdot p_{unk} \cdot w_{unk} \quad (16)$$

$$m(\text{untrust}) = w \cdot (1 - r) \quad (17)$$

These discounted belief masses are then fused iteratively using Dempster's combination rule, which aggregates belief from multiple sources while adjusting for conflicts among them. Specifically, for any two distributions m_1 and m_2, their combined belief mass for each subset $A \subseteq \{H1, H2, H1\&H2\}$ is computed by identifying the intersections of their focal elements and normalizing by the total conflict. The result is a new fused distribution over the three hypotheses and a residual untrustworthy mass that captures irreconcilable conflicts or low confidence.

In the context of this paper, I apply the ER framework to perform three fusions and constructed a three-level MAKER model, and uncertain outcomes for each document segment, with the untrustworthy component retained separately to reflect information noise or ambiguity. This evidential reasoning process not only preserves the interpretability of individual sources but also allows for fine-grained control of their influence through calibrated parameters. In this study, I adopt empirically grounded default values for the weights and reliability scores based on citation-based credibility and cross-validation, ensuring both transparency and empirical validity. The ER framework thus serves as the backbone for each level of the proposed model, enabling robust and interpretable sentiment aggregation across dictionaries, roles and indicators.

Building upon the evidential reasoning framework, I implement a three-level MAKER architecture to operationalize sentiment extraction, structural aggregation, and stock prediction. In the first level, I address the challenge of heterogeneous information by extracting sentiment signals from five key textual components: the presentation, question, and answer sections of the earnings call, as well as the corresponding 10-K and 10-Q filings. For each component, I apply four complementary financial dictionaries to capture sentiment. Each dictionary acts as an independent source of evidence, and their outputs are fused using the MAKER model. Rather than relying on simple word ratios, I adopt a sentiment coverage adjustment method that considers both the frequency and contextual weight of sentiment words

within the entire text, producing interpretable sentiment probabilities for each document segment.

In the second level, I further aggregate the three conference call sections into a unified sentiment profile. These sections differ in both linguistic tone and perceived informativeness. Analyst questions often attract the most attention due to their directness, while managerial answers may be spontaneous and less reliable. The presentation section, on the other hand, tends to be more standardized and carefully crafted. To account for these differences, I assign differential weights and reliabilities to each section based on their expected influence and credibility. These are used as parameters within the MAKER model to produce a consolidated sentiment distribution for the entire call, while also retaining a separate probability mass to reflect residual untrustworthiness.

The third level introduces a behavioural dimension. I compare the overall sentiment of the current call with a reference point constructed from recent 10-K and 10-Q reports. This difference captures how investors might perceive the direction and tone of new information relative to prior expectations. Drawing on insights from prospect theory, I model how such deviations are psychologically interpreted—whether perceived as gains or losses, and how their likelihood is subjectively weighted. Finally, I combine these behavioural evaluations with structured firm-level financial data and use evidential reasoning to produce a prediction of stock price movement following the call. This process generates a final probability distribution regarding whether stock prices will rise or fall, while maintaining the model's complete interpretability.

Together, these three levels form a coherent, evidence-based framework for modeling how investors process, evaluate, and act upon qualitative disclosures. The MAKER structure enables the model to handle both heterogeneous and uncertain information, while also capturing cognitive distortions and behavioural tendencies in financial decision-making.

3.3 Prospect theory

In the application of prospect theory in this study, I introduce the sentiment of annual and quarterly reports as the reference point. Annual and quarterly reports all have a specific section that reflects the content of discussions between analysts and managers, which has a certain degree of sentiment, especially annual reports, which have always been the focus of textual analysis. And based on the outputs from the first-level and second-level MAKER models, I derive a comprehensive sentiment distribution for each earnings call. I then compute the deviation between this call-level sentiment and the sentiment extracted from the firm's most recent 10-K and 10-Q filings. This deviation serves as the input to a behavioural evaluation process grounded in prospect theory. After assigning the uncertain sentiments from the conference call to positive and negative sentiments according to investor preferences, I apply a value function to capture the asymmetric sensitivity to gains and losses and a probability weighting function to model investors' tendency to distort objective probabilities. Through this behavioural transformation, the model links textual sentiment dynamics to investor perceptions, allowing for a nuanced simulation of how shifts in corporate tone influence subsequent market reactions.

$$v(intens_{i,j}) = intens_{i,j}{}^{\alpha}, intens_{i,j} > 0 \tag{18}$$

$$v(intens_{i,j}) = -\lambda(-intens_{i,j})^{\beta}, intens_{i,j} < 0 \tag{19}$$

For firm i in quarter t, the value of every type of words can be calculated in each part by using value function. $intens_{i,j}$ represents the intensity of words. Where, $0< \alpha < 1$, $0 < \beta < 1, \lambda > 0$. Then, weight functions are applied to analyze the percentages of every sentiment.. For firm i in quarter t, weight functions are defined by:

$$w^{+}(p_{i,t}) = \exp(-\beta^{+} \cdot (-lnp_{i,t})^{\alpha^{+}}) \tag{20}$$

$$w^{+}(p_{i,t}) = \exp(-\beta^{-} \cdot (-lnp_{i,t})^{\alpha^{-}}) \tag{21}$$

Where $w^{+}(0) = w^{-}(0) = 0,\ \ w^{+}(1) = w^{-}(1) = 1$. Finally, the value function can be weighted by the weight function to obtain evaluation of conference calls $PV_{i,t}$.

4. Results

For the first stage of empirical analysis, I have completed all preprocessing and integration steps. To conduct preliminary model evaluation, I draw a stratified subsample of 952 calls from the full dataset. Specifically, I randomly select approximately 50 samples per year from 2004 to 2024 to ensure balanced temporal coverage. In the earlier years of the sample period, the total number of earnings calls per year was fewer than 50. To preserve as much temporal information as possible, I retain all available calls from those years in the subsample, rather than applying random sampling. This subsample enables me to validate the feasibility of the proposed methodology and conduct early-stage parameter tuning. Full-sample estimation and robustness testing will be implemented in the subsequent stage after finalizing the model's structural settings.

To implement the parameter estimation procedure, I divide the subsample into training and testing sets based on calendar years. Specifically, the data from 2004 to 2020 is designated as the training set for parameter optimization. The remaining data from 2021 to 2024 serves as the testing set to evaluate the model's generalization performance. This temporal split ensures a strict forward-looking structure, preventing data leakage and maintaining the causal direction required for behavioral inference. To estimate all 122 parameters, I employ Particle Swarm Optimization (PSO), a population-based metaheuristic that allows for constrained, structured, and interpretable search. This method is well-suited for the model's high-dimensional, constrained, and non-convex parameter space. By initializing parameters with behaviorally interpretable values and enforcing theoretical constraints, PSO can perform efficient global searches while incorporating expert knowledge. This stage focuses on validating the proposed estimation strategy and assessing the model's empirical viability before scaling to the full dataset.

Table 1. Prediction Performance of $BHRET\_0\_1$ (Initial Parameters)

Accuracy	F1 Score	AUC	Precision	Recall	Balanced Accuracy	Log Loss
0.5651	0.5500	0.5914	0.5418	0.5585	0.5648	0.6802

To evaluate the performance and interpretability of the proposed MAKER framework, I conduct a two-stage empirical assessment. The prediction target is the one-day cumulative abnormal return (BHRET_0_1), a widely used metric to capture the immediate market reaction following earnings disclosures. In the first stage, I test the predictive power of the MAKER model using manually specified initial parameters. Most of these parameters are interpretable, derived from expert knowledge. Despite no optimization, the model achieves an accuracy of 56.51%, an F1 score of 0.5500, and an AUC of 0.5914. These performance metrics are well above the chance level (50%) and illustrate that the structural design of the MAKER model captures meaningful decision-relevant information. This result highlights the model's strength in interpretability and theoretical alignment, even prior to any data-driven calibration.

Figure 1. Receiver Operating Characteristic (ROC) Curve (Initial Parameters)

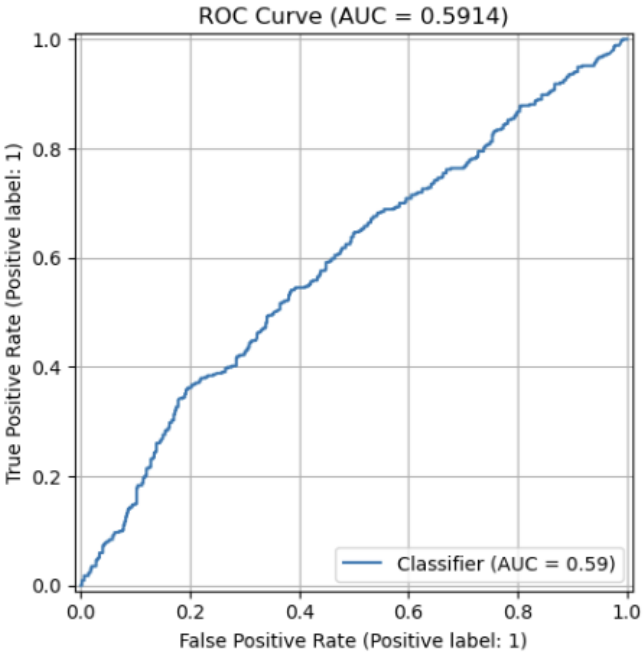


Figure 2. Confusion Matrix of Prediction Results (Initial Parameters)

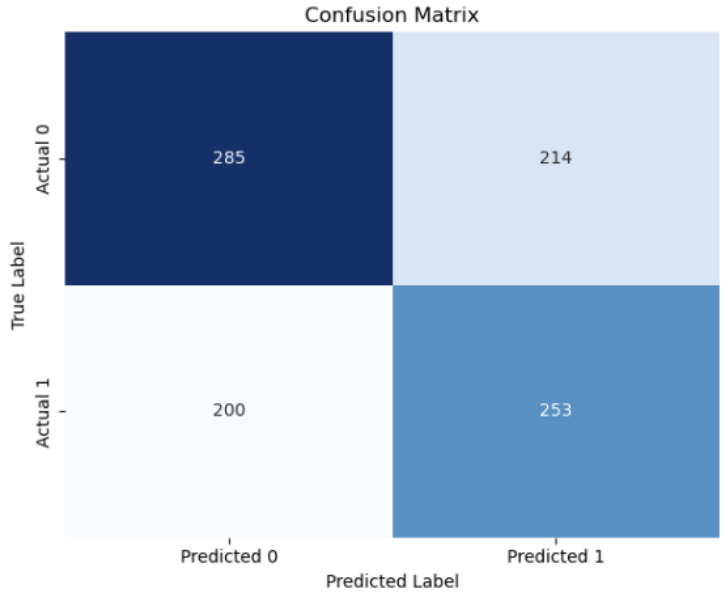


Figure 3. Predicted Probabilities vs. True Labels (Initial Parameters)

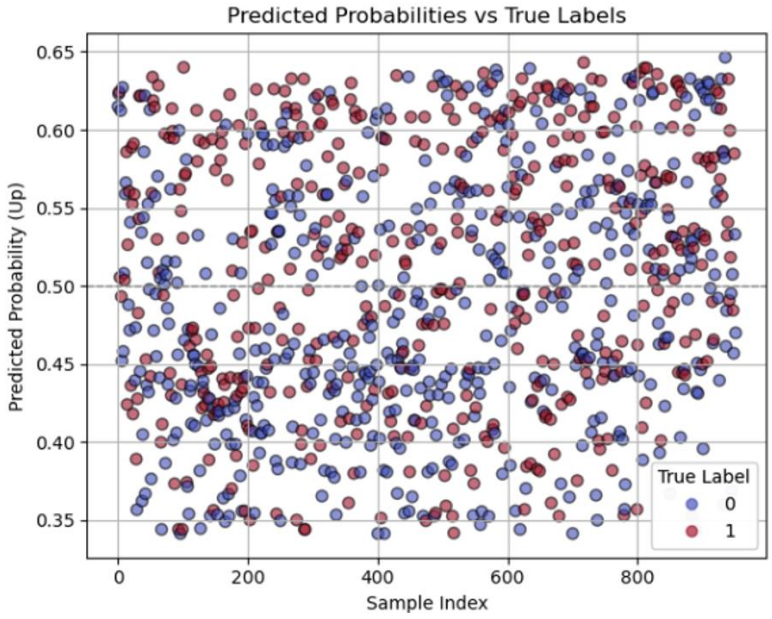


Figure 1 presents the Receiver Operating Characteristic (ROC) curve, which yields an area under the curve (AUC) of 0.5914. This performance, though not high, is notably better than random guessing and demonstrates that even without optimization, the MAKER model has the capacity to distinguish between positive and negative abnormal returns. Figure 2 reports the confusion matrix corresponding to the initial

parameter predictions. The model correctly classifies 285 negative-return cases and 253 positive-return cases, indicating a relatively balanced ability to capture both types of return directions. Figure 3 visualizes the predicted probabilities against the true labels across all samples. Although the predicted probabilities are moderately dispersed, a degree of separation between red (positive returns) and blue (negative returns) dots is observable, especially in the upper and lower probability ranges. Collectively, these figures provide visual evidence that the initial parameter configuration produces a meaningful predictive signal and offers a credible foundation for subsequent optimization.

Table 2. Prediction Performance of $BHRET\_0\_1$ (After Optimization)

Accuracy	F1 Score	AUC	Precision	Recall	Balanced Accuracy	Log Loss
0.6134	0.5721	0.6398	0.6044	0.5430	0.6102	0.6621

In the second stage, I perform parameter optimization using particle swarm optimization (PSO). After optimization, the model's performance improves notably: accuracy rises to 61.34%, F1 score increases to 0.5721, and AUC reaches 0.6398. These improvements validate the necessity and effectiveness of optimization. Importantly, the optimized parameters still respect the original interpretability structure, such as ordered reliability among textual components and the behavioural shapes of prospect theory functions, suggesting that optimization enhances rather than overrides the behavioural design.

Figure 4. Receiver Operating Characteristic (ROC) Curve (After Optimization)

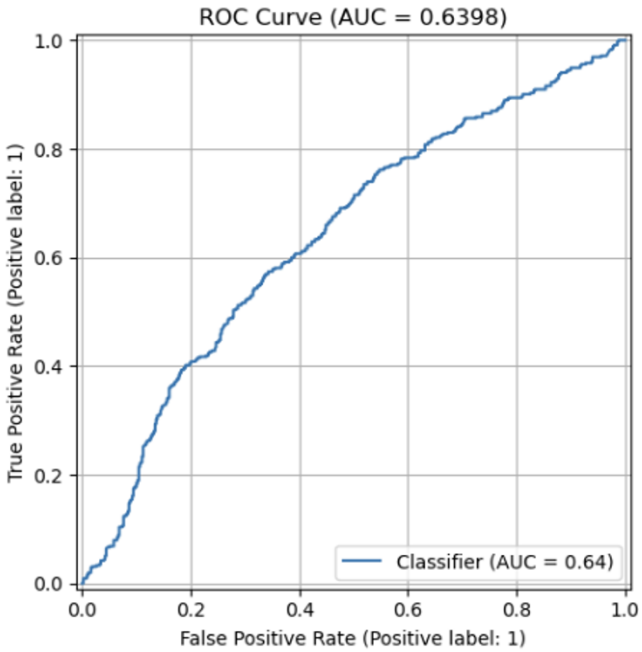


Figure 5. Confusion Matrix of Prediction Results (After Optimization)

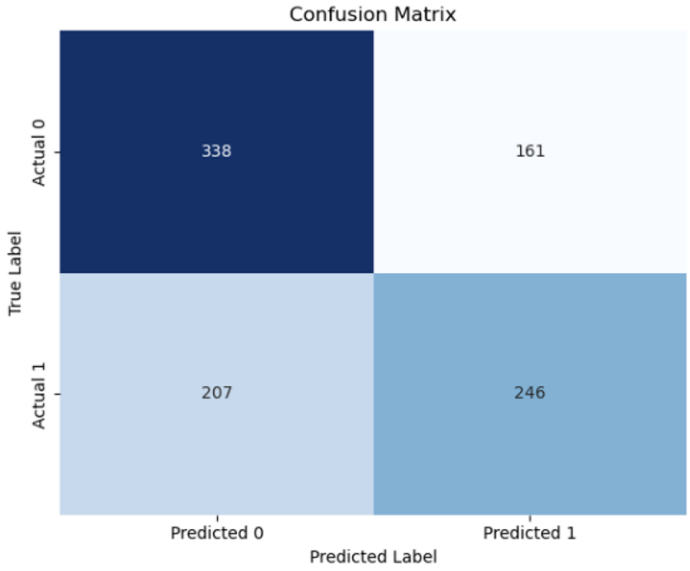


Figure 6. Predicted Probabilities vs. True Labels (After Optimization)

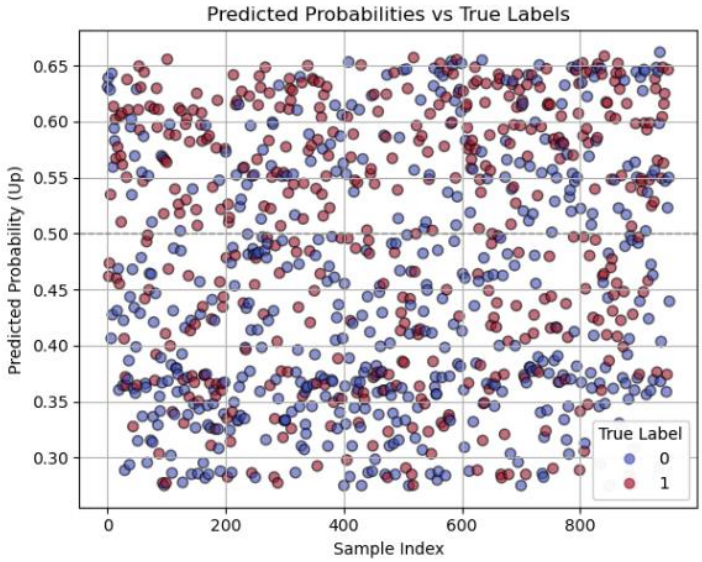


Figure 4 displays the Receiver Operating Characteristic (ROC) curve after optimization. Compared to the initial configuration, the AUC increases from 0.5914 to 0.6398, indicating a more reliable distinction between positive and negative returns across all thresholds. Figure 5 presents the updated confusion matrix, where the model correctly identifies 338 negative-return samples and 246 positive-return samples. Notably, the number of false positives drops from 214 to 161, and true negatives increase from 285 to 338, demonstrating an improved ability to minimize Type I errors while still maintaining sensitivity. These gains are consistent with the rise in both precision (from 54.18% to 60.44%) and recall (from 55.85% to 54.30%), achieving a more balanced predictive performance. Figure 6 plots the predicted probabilities versus true labels after optimization. Compared to the pre-optimization scatter, the distribution of red (positive) and blue (negative) dots becomes more stratified along the y-axis, especially in the higher (0.55–0.65) and lower (0.30–0.40) probability bands. This visual evidence suggests a clearer separation in predicted confidence, reinforcing the model's improved calibration. Together, these figures provide strong support for the MAKER model's capacity to deliver both interpretability and predictive power, with optimization refining rather than distorting its foundational structure.

It is important to note that this evaluation is based on a pilot sample of only 952 calls. The full dataset comprises 62,257 earnings calls spanning two decades, and will be used in the final deployment of the MAKER model. Given the robustness shown in this small sample, I expect performance to improve substantially with the larger sample, enabling better generalization. Overall, this empirical evaluation provides evidence for three key points: first, the initial interpretable design of the MAKER model delivers meaningful predictions without overfitting; second, PSO optimization meaningfully enhances predictive accuracy while preserving interpretability; and third, the model is scalable and ready for full-sample deployment in large-scale financial forecasting tasks.

5. Research conclusion

In recent years, the rapid development of artificial intelligence technology has brought new research perspectives to the analysis of textual characteristics in earnings conference calls. In previous studies, the predictive model of conference calls sentimental words for stock returns did not fully digging out all textual information. In this study, I construct a multi-stage research framework aimed at understanding investor behaviour and improving stock price prediction through interpretable artificial intelligence, which develop a fully interpretable three-level evidential reasoning model.

Empirical evaluation based on a random sample of 952 instances demonstrates that even with manually defined initial parameters—without any machine learning optimization—the model achieves reasonably strong predictive performance. These results validate the behavioural structure and interpretability of the model, as the predictions are driven by transparent and psychologically grounded mechanisms. Furthermore, after applying Particle Swarm Optimization (PSO) to optimise the parameters, the model exhibits significant improvement in predictive accuracy, confirming that the interpretable framework not only preserves behavioural transparency but also holds substantial potential for empirical performance

enhancement.

The insights obtained at this stage provide a robust foundation for future research. In subsequent work, I will investigate how to integrate this interpretable framework with more complex yet less transparent models, such as BERT enhanced with SHAP explanations and ChatGPT-based architectures. The goal is to develop an ensemble learning model that effectively balances interpretability and predictive performance. Through this approach, I aim to deepen our understanding of how investors process information under uncertainty and offer more rigorous support for applications in behavioural finance.

Reference

Andrew, C. et al. (2021) 'Which Buy-Side Institutions Participate in Public Earnings Conference Calls? Implications for Capital Markets and Sell-Side Coverage', Journal of Corporate Finance, 68, pp. 101-964

Blau, B M., Delisle, J R. and Price, S M. (2015) 'Do Sophisticated Investors Interpret Earnings Conference Call Tone Differently than Investors at Large? Evidence from Short Sales', Journal of Corporate Finance, 31, pp. 203-219.

Bochkay, K., Hales, J. and Chava, S. (2020) 'Hyperbole or Reality? Investor Response to Extreme Language in Earnings Conference Calls', ACCOUNTING REVIEW, 95(2), pp. 31–60. Available at: https://doi.org/10.2308/accr-52507.

Bowen, R T,. Davis, A K. and Matsumoto, D A. (2002) 'Do Conference Calls Affect Analysts' Forecasts?', The Accounting review, 77(2), pp. 285–316.

Bracha, A. (2020) 'Investment Decisions and Negative Interest Rates', Management Science, 66(11), pp. 5316–5340. Available at: https://doi.org/10.1287/mnsc.2019.3464.

Bracha, A. and Brown, D.J. (2012) 'Affective decision making: A theory of optimism bias', Games and economic behavior, 75(1), pp. 67–80. Available at: https://doi.org/10.1016/j.geb.2011.11.004.

Brochet, Francois, Kolev, Kahn, Lerman and Alina (2018), Information Transfer and Conference Calls. Review of Accounting Studies.

Brown, S., Hillegeist, S A. and Lo, K. (2004) 'Conference Calls and Information Asymmetry', Journal of Accounting and Economics, 37(3), pp. 343-366.

Bushee, B J., Matsumoto, D A. and Miller, G S. (2003) 'Open versus Closed Conference Calls: The Determinants and Effects of Broadening Access to Disclosure', Journal of Accounting & Economics, 34(1–3), pp. 149–180.

Cabrera-Paniagua, D. et al. (2015) 'Decision making system for stock exchange market using artificial emotions', Expert systems with applications, 42(20), pp. 7070–7083. Available at: https://doi.org/10.1016/j.eswa.2015.05.004.

Call, A.C. et al. (2023) 'Managers' use of humor on public earnings conference calls', Review of accounting studies [Preprint]. Available at: https://doi.org/10.1007/s11142-023-09764-x.

Campitelli, G. and Gobet, F. (2010) 'Herbert Simon's Decision making Approach: Investigation of Cognitive Processes in Experts', Review of General Psychology, 14(4), pp. 354–364. Available at: https://doi.org/10.1037/a0021256.

Cavagnaro, D.R. et al. (2013) 'Discriminating among probability weighting functions using adaptive design optimization', Journal of risk and uncertainty, 47(3), pp. 255–289. Available at: https://doi.org/10.1007/s11166-013-9179-3.

Chai, J. and Ngai, E.W. (2020) 'The variable precision method for elicitation of probability weighting functions', Decision Support Systems, 128, pp. 113166-. Available at: https://doi.org/10.1016/j.dss.2019.113166.

Ciccarone, G., Giuli, F. and Marchetti, E. (2020) 'Prospect Theory and sentiment-driven fluctuations', The B.E. journal of macroeconomics, 20(1). Available at: https://doi.org/10.1515/bejm-2017-0118.

Davis, A K. et al. (2015) 'The Effect of Manager-Specific Optimism on the Tone of Earnings Conference Calls', Review of Accounting Studies, 20(2), pp. 639-673.

De Giorgi, E., Hens, T. and Rieger, M.O. (2010) 'Financial market equilibria with cumulative prospect theory', Journal of mathematical economics, 46(5), pp. 633–651. Available at: https://doi.org/10.1016/j.jmateco.2010.06.001.

Doran, J S., Peterson, D R. and Price, M K. (2012) 'Earnings Conference Call Content and Stock Price: The Case of REITs', The Journal of Real Estate Finance and Economics, 45(2), pp. 402–434.

Frankel, R M. et al. (1997) 'An Empirical Evaluation of Conference Calls as a Voluntary Disclosure Medium', Journal of Accounting Research, 30(2), pp. 21-23.

Fu, X., Wu, X. and Zhang, Z. (2021) 'The Information Role of Earnings Conference Call Tone: Evidence from Stock Price Crash Risk', JOURNAL OF BUSINESS ETHICS, 173(3), pp. 643–660.

Garcés, M. and Finkel, L. (2019) 'Emotional Theory of Rationality', Frontiers in integrative neuroscience, 13, pp. 11–11. Available at: https://doi.org/10.3389/fnint.2019.00011.

Garcia, D., Hu, X. and Rohrer, M. (2023) 'The colour of finance words', JOURNAL OF FINANCIAL ECONOMICS, 147(3), pp. 525–549. Available at:

https://doi.org/10.1016/j.jfineco.2022.11.006.

Gazioglu, S. and Caliskan, N. (2011) 'Cumulative prospect theory challenges traditional expected utility theory', Applied financial economics, 21(21), pp. 1581–1586. Available at: https://doi.org/10.1080/09603107.2011.583393.

Glöckner, A. and Pachur, T. (2012) 'Cognitive models of risky choice: Parameter stability and predictive accuracy of prospect theory', Cognition, 123(1), pp. 21–32. Available at: https://doi.org/10.1016/j.cognition.2011.12.002.

Gonzalez, R. and Wu, G. (1999) 'On the Shape of the Probability Weighting Function', Cognitive psychology, 38(1), pp. 129–166. Available at: https://doi.org/10.1006/cogp.1998.0710.

Grayot, J.D. (2020) 'Dual Process Theories in Behavioral Economics and Neuroeconomics: a Critical Review', Review of philosophy and psychology, 11(1), pp. 105–136. Available at: https://doi.org/10.1007/s13164-019-00446-9.

Huang, A H, Lehavy, R, Zang, A Y and Zheng, R (2017), Analyst Information Discovery and Interpretation Roles: A Topic Modeling Approach. Management Science: mnsc.2017.2751.

Huang, A H., Zang, A Y. and Rong, Z. (2014) 'Evidence on the Information Content of Text in Analyst Reports', Accounting Review, 89(6), pp. 2151–2180.

Jancenelle, V.E., Storrud-Barnes, S.F. and Iaquinto, A. (2019) 'Making investors feel good during earnings conference calls: The effect of warm-glow rhetoric', JOURNAL OF GENERAL MANAGEMENT, 44(2), pp. 63–72. Available at: https://doi.org/10.1177/0306307018813759.

Jason, V. et al. (2018),' Manager-Analyst Conversations in Earnings Conference Calls', Review of Accounting Studies, 23(4), pp. 1315-1354.

Juechems, K. et al. (2021) 'Optimal utility and probability functions for agents with finite computational precision', Proceedings of the National Academy of Sciences - PNAS, 118(2), pp. 1-. Available at: https://doi.org/10.1073/pnas.2002232118.

Karmarkar, U.S. (1978) 'Subjectively weighted utility: A descriptive extension of the expected utility model', Organizational behavior and human performance, 21(1), pp. 61–72. Available at: https://doi.org/10.1016/0030-5073(78)90039-9.

Kilka, M. and Weber, M. (2001) ‘What Determines the Shape of the Probability Weighting Function Under Uncertainty?’, Management science, 47(12), pp. 1712–1726. Available at: https://doi.org/10.1287/mnsc.47.12.1712.10239.

Kirby, K.N. (2011) ‘An Empirical Assessment of the Form of Utility Functions’, Journal of experimental psychology. Learning, memory, and cognition, 37(2), pp. 461–476. Available at: https://doi.org/10.1037/a0021968.

Kirshner, S.N. and Shao, L. (2019) ‘The overconfident and optimistic price-setting newsvendor’, European journal of operational research, 277(1), pp. 166–173. Available at: https://doi.org/10.1016/j.ejor.2019.02.023.

Krawczyk, M.W. (2015) ‘Probability weighting in different domains: The role of affect, fungibility, and stakes’, Journal of economic psychology, 51, pp. 1–15. Available at: https://doi.org/10.1016/j.joep.2015.06.006.

Lee, J. (2016) ‘Can Investors Detect Managers’ Lack of Spontaneity? Adherence to Predetermined Scripts during Earnings Conference Calls’, Accounting Review, 91(1), pp. 229–250.

Loughran, T. and Mcdonald, B. (2011) ‘When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks’, Journal of Finance, 66(1), pp. 35-36.

Matsumoto, D., Pronk, M. and Roelofsen, E. (2011) ‘What Makes Conference Calls Useful? The Information Content of Managers’ Presentations and Analysts’ Discussion Sessions’, Accounting Review, 86(4), pp. 1383–1414.

Mello, F.L. de and Souza, S.A. de (2021) ‘Decision Maker Profiling Using Their Mental Behavior Pattern’, Frontiers in psychology, 12, pp. 667255–667255. Available at: https://doi.org/10.3389/fpsyg.2021.667255.

Mukherjee, K. (2010) ‘A Dual System Model of Preferences Under Risk’, Psychological review, 117(1), pp. 243–255. Available at: https://doi.org/10.1037/a0017884.

Mukherjee, K. (2011) ‘Thinking styles and risky decision-making: Further exploration of the affect-probability weighting link’, Journal of behavioral decision making, 24(5), pp. 443–455. Available at: https://doi.org/10.1002/bdm.700.

Neumann, J. Von and Morgenstern, O. (1944) The Theory of Games and Economic

Behaviour. Princeton University Press.

Nilsson, H., Rieskamp, J. and Wagenmakers, E.-J. (2011) 'Hierarchical Bayesian parameter estimation for cumulative prospect theory', Journal of mathematical psychology, 55(1), pp. 84–93. Available at: https://doi.org/10.1016/j.jmp.2010.08.006.

Prelec, D. (1998) 'The Probability Weighting Function', Econometrica, 66(3), pp. 497–527. Available at: https://doi.org/10.2307/2998573.

Price, S M. et al. (2012) 'Earnings Conference Calls and Stock Returns: The Incremental Informativeness of Textual Tone', Journal of banking & finance, 36(4), pp. 992–1011.

Price, S.M. et al. (2012) 'Earnings conference calls and stock returns: The incremental informativeness of textual tone', JOURNAL OF BANKING & FINANCE, 36(4), pp. 992–1011. Available at: https://doi.org/10.1016/j.jbankfin.2011.10.013.

Sahlin, N.-E., Wallin, A. and Persson, J. (2010) 'Decision Science: From Ramsey to Dual Process Theories', Synthese (Dordrecht), 172(1), pp. 129–143. Available at: https://doi.org/10.1007/s11229-009-9472-5.

Steele, K.S. (2010) 'What are the minimal requirements of rational choice? Arguments from the sequential-decision setting', Theory and decision, 68(4), pp. 463–487. Available at: https://doi.org/10.1007/s11238-009-9145-3.

Suslava, K. (2021) '"Stiff Business Headwinds and Uncharted Economic Waters": The Use of Euphemisms in Earnings Conference Calls', MANAGEMENT SCIENCE, 67(11), pp. 7184–7213. Available at: https://doi.org/10.1287/mnsc.2020.3826.

Tversky, A. and Kahneman, D. (1992) 'Advances in prospect theory: Cumulative representation of uncertainty', Journal of Risk and Uncertainty, 5(4), pp. 297–323.

Tversky, K. A. (1979) 'Prospect Theory: An Analysis of Decision under Risk', Econometrica, 47(2), pp. 263–291.

van de Kuilen, G. and Wakker, P.P. (2011) 'The Midweight Method to Measure Attitudes Toward Risk and Ambiguity', Management science, 57(3), pp. 582–598. Available at: https://doi.org/10.1287/mnsc.1100.1282.

Wu, G. and Gonzalez, R. (1996) 'Curvature of the Probability Weighting Function', Management science, 42(12), pp. 1676–1690. Available at:

https://doi.org/10.1287/mnsc.42.12.1676.

Yang, J.B. (1999) ‘Rule and utility based evidential reasoning approach for multiattribute decision analysis under uncertainties’, European Journal of Operational Research, 131(1), pp. 31–61. Available at: https://doi.org/10.1016/S0377-2217(99)00441-5.

Yang, J.B. and Xu, D.L. (2025) ‘Maximum Likelihood Evidential Reasoning’, Artificial Intelligence, 340(January), p. 104289. Available at: https://doi.org/10.1016/j.artint.2025.104289.