Green AI in Action:

Strategic Model Selection for Ensembles in Production

Nienke Nijkamp

Delft University of Technology

TU Delft

Green AI in Action:

Strategic Model Selection for Ensembles in Production

by

Nienke Nijkamp

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Thursday April 11, 2024 at 10:00 AM.

Thesis advisor:	Arie van Deursen	*TU Delft*
Daily supervisor:	Luís Miranda da Cruz	*TU Delft*
Daily co-supervisor:	June Sallou	*TU Delft*
	Niels van der Heijden	*Deloitte*
Thesis committee:	Luciano Cavalcante Siebert	*TU Delft*

Student Number:	4659872
Programme:	MSc Computer Science, Delft
Project Duration:	September, 2023 - April, 2024

TU Delft

Preface

This thesis examines how model selection strategies can impact the accuracy and efficiency of a live AI ensemble system. To achieve this, we collaborated with Deloitte NL to study the AI system DocQMiner. This thesis is conducted to obtain the degree of Master of Science in Computer Science.

Having spent a year working at Deloitte, I was excited to partner with them for my research. I'm profoundly thankful for the chance to merge my interest in Sustainable AI with a real-world industry context. This journey afforded me a practical understanding of the complexities and challenges inherent in AI systems within professional service settings.

I extend my sincerest gratitude to my thesis advisor, Arie van Deursen, and my daily supervisors, Luís Miranda da Cruz, and June Sallou, whose guidance was instrumental in shaping this research. Their expertise and patience were invaluable throughout this process. My gratitude also extends to Niels van der Heijden from Deloitte, who provided invaluable industry insights and always made time to share his knowledge during the many brainstorming sessions.

Furthermore, I would like to thank my colleagues at Deloitte, especially the DocQMiner and Platform team, for their patience and willingness to answer my countless questions. Their valuable insights and viewpoints significantly influenced the execution of this research.

To my family and friends, your constant support and encouragement have been the cornerstone of this thesis and my entire academic journey. I am incredibly thankful for the comforting spaces you provided, allowing me to focus and thrive during this time.

As I present this thesis for public defense, I reflect on the insights and knowledge gained by this research. The journey has been challenging and transformative, and I share these findings with a sense of accomplishment and anticipation.

Nienke Nijkamp
Amsterdam, April 2024

Abstract

Integrating Artificial Intelligence (AI) into software systems has significantly enhanced their capabilities while escalating energy demands. Ensemble learning, combining predictions from multiple models to form a single prediction, intensifies this problem due to cumulative energy consumption.

This research presents a novel approach to model selection that addresses the challenge of balancing the accuracy of AI models with their energy consumption in a live AI ensemble system. We explore how reducing the number of models or improving the efficiency of model usage within an ensemble during inference can reduce energy demands without substantially sacrificing accuracy.

This study introduces and evaluates two model selection strategies, *Static* and *Dynamic*, for optimizing ensemble learning systems' performance while minimizing energy usage. Our results demonstrate that the *Static* strategy improves the F1 score beyond the baseline, reducing average energy usage from 100% from the full ensemble to 62%. The *Dynamic* strategy further enhances F1 scores, while using on average 76% compared to 100% of the full ensemble.

Moreover, we propose an approach that balances accuracy with resource consumption, significantly reducing energy usage without substantially impacting accuracy. This method decreased the average energy usage of the *Static* strategy from approximately 62% to 14%, and for the *Dynamic* strategy, from around 76% to 57%.

This research aligns with the principles of Green AI by demonstrating that strategic model selection can achieve a balance between accuracy and computational efficiency.

Our field study of Green AI using an operational AI system developed by a large professional services provider shows the practical applicability of adopting energy-conscious model selection strategies in live production environments.

Contents

1 Introduction

Over recent years, the integration of Artificial Intelligence (AI) in modern software systems has been growing [25]. However, the incorporation of AI, and specifically AI systems, has led to a significant rise in resource demands, including the energy required for training and deploying these models [51, 2]. Ensemble learning, a method that combines multiple models to create a more effective solution, while highly effective, tends to intensify this energy consumption problem due to the cumulative energy requirements of the individual models [37]. Our study is geared towards reducing energy consumption while maintaining accuracy in a live AI system.

1.1. Green AI

Over recent years, the integration of Artificial Intelligence (AI) in modern software systems has been growing [25]. Consequently, the incorporation of AI, specifically AI systems, has led to a significant rise in resource demands, including the energy required for training, deploying, and inferencing these models [51, 2, 11]. To illustrate, a ChatGPT-like application handling 11 million requests per hour is estimated to emit 12,800 tons of CO_2 annually, making inference 25 times more carbon-intensive than training GPT-3 [11].

The focus of the AI research community has predominantly been on improving the accuracy of AI models, overlooking the significant energy costs associated with them. AI's rising environmental and financial cost has led to a pressing need for a more balanced approach to AI development that considers accuracy and energy efficiency [51]. The emerging field of Green AI addresses this gap, promoting a favorable trade-off between efficiency and accuracy [47].

There exist various benefits to implementing Green AI, such as:

- **Reduced environmental impact**: Green AI aims to minimize the energy consumption and carbon emissions associated with AI models, thereby mitigating their environmental impact [51, 11]. This is critical in the context of global efforts to combat climate change and reduce greenhouse gas emissions.
- **Cost savings**: By optimizing AI models for energy efficiency, organizations can achieve significant cost savings, particularly in terms of reduced energy bills and lower infrastructure expenses[2]. This makes AI more accessible and affordable, especially for smaller entities and startups.
- **Enhanced performance**: Green AI encourages the development of lightweight and efficient models, which can often perform better in resource-constrained environments. This leads to faster inference times and improved usability of AI applications on edge devices[27].
- **Social responsibility**: Adopting Green AI practices demonstrates a commitment to social responsibility and ethical AI development[13]. It aligns with the broader goals of sustainable development and responsible technology use.

A significant gap remains in implementing Green AI principles within the industry [53]. Despite the growing awareness of the environmental impact of AI and the potential benefits of Green AI, the adoption of sustainable practices in AI development is still limited. Challenges such as the lack of standardized metrics for measuring the resource consumption of AI models, the absence of incentives for developing energy-efficient algorithms, and the need for a cultural shift towards prioritizing sustainability in AI research and development are some of the barriers to widespread implementation [47, 33, 57].

1.2. Ensemble Learning

Ensemble learning is a powerful technique in the field of machine learning that involves combining multiple models to improve the overall performance of a system [59]. The concept of ensemble learning can be traced back to the early 1990s, with significant contributions from researchers such as Hansen and Salamon [28] and Breiman [5]. The primary motivation behind ensemble learning is the recognition that different models may capture different aspects of the data, and by aggregating their predictions, one can achieve a more robust and accurate outcome.

There are various types of ensemble methods, each with its unique approach to combining models:

- **Bagging (Bootstrap Aggregating)**: Introduced by Breiman [5], bagging involves training multiple models on different subsets of the training data (sampled with replacement) and then averaging their predictions. Random Forest [6] is a well-known example of a bagging ensemble.

- **Boosting**: Boosting algorithms, such as AdaBoost [22] and Gradient Boosting [23], sequentially train models, with each model focusing on the instances that were misclassified by the previous models. The final prediction is a weighted sum of the individual models' predictions.

- **Stacking**: Stacking involves training multiple models on the same data and then using a meta-model to combine their predictions [56]. The meta-model is trained on the predictions of the base models, allowing it to learn how to best combine their outputs.

Popular applications of ensemble learning are speech recognition [18, 36] and classification [52, 42], image classification [54, 35, 31] and forecasting [44, 38, 7, 49]. Additionally, ensemble methods have been successful in winning numerous machine learning competitions, such as the Netflix Prize [3] and Kaggle competitions [10].

The relevance of ensemble learning extends beyond its predictive capabilities. In the context of Green AI, ensemble methods pose both challenges and opportunities. While ensembles can be more resource-intensive due to the use of multiple models [37, 15], they also offer a potential pathway for balancing accuracy and energy efficiency. By selectively combining models or employing energy-efficient ensemble techniques, one can achieve high-performance AI systems that are also environmentally sustainable [37, 55].

1.3. Deloitte

Deloitte began as an accountancy firm but has since expanded into a broad range of professional services. Founded by William Welch Deloitte in 1845 in London, Deloitte initially focused on audit and financial advisory services. Over time, the firm has significantly broadened its service offerings to include management consulting, risk advisory, tax, and other related services. Today, Deloitte is one of the "Big Four" professional services networks and offers a wide spectrum of services to clients across various industries around the world.

As the largest professional services network in the world by revenue and number of professionals, Deloitte offers a multifaceted approach to risk management and mitigation. Leveraging its expansive global footprint and cross-industry experience, Deloitte assists organizations in navigating digital risks, from cybersecurity threats to compliance issues.

1.3.1. DocQMiner

DocQMiner is a tool developed by Deloitte's Digital Risk Solutions (DRS) team. The DRS team consists of around 65 professionals who assist businesses in effectively identifying, assessing, and mitigating various types of risks using assets, ensuring resilience and compliance in the corporate landscape.

DocQMiner was developed upon a client request in 2018 when thousands of documents had to be processed within three months. Development continued, and the tool is currently used by many different customers in various domains, such as the public sector, life sciences, and financial services.

1.4. Research question

Our research explores the intersection of Green AI and ensemble learning in a production environment, a domain that has yet to be thoroughly investigated. Simply running all models every time is not an efficient strategy [60]. The challenge is finding a more intelligent approach for running inference with ensemble learning [55], reducing energy consumption while maintaining or improving accuracy.

Our research explores the intersection of Green AI and ensemble learning in a production environment, a domain that has yet to be thoroughly investigated. Simply running all models every time is not an efficient strategy [60]. The challenge is finding a more intelligent approach for running inference with ensemble learning [55], reducing energy consumption while maintaining or improving accuracy.

In this research, we propose a solution that involves a selective approach to using models within an ensemble. The core of this approach is the concept of *model selection strategies*, which refers to various methods of selecting specific subsets of models for individual tasks rather than using the complete set of models for every task [8]. Our goal in implementing these model selection strategies is to balance achieving accuracy and managing computational costs. This goal leads to the following research question.

Research Question: What are the impacts of implementing model selection strategies on the accuracy and energy usage of ensemble learning systems?

Two model selection strategies, *Static* and *Dynamic*, are investigated for optimizing model performance. Both strategies start by evaluating all subsets of the entire ensemble on a validation set, selecting the combination with the highest accuracy. *Static* selection identifies the best overall model selection, while *Dynamic* selection chooses the best selection per specific property within the domain.

Additionally, we consider the computational cost per model by employing a metric that discounts accuracy with energy consumption. This method is incorporated into both the *Static* and *Dynamic* selection strategies, ensuring a balanced consideration of performance and resource efficiency.

This research provides practical insights into making model ensembles more efficient by examining and implementing our model selection strategies within an ensemble learning context. Our contributions to the field of AI in software systems using ensemble learning are the following:

1. A detailed evaluation of *Static* and *Dynamic* model selection strategies in a production environment.
2. An approach to enhance these strategies by incorporating energy usage metrics, significantly lowering energy consumption.

These contributions demonstrate that *model selection strategies* not only significantly reduce resource consumption but also have the potential to maintain or increase the accuracy of the AI system.

Furthermore, these findings highlight the practicality and necessity of integrating Green AI principles into AI development, working towards more sustainable and efficient AI applications in the industry[1].

1.5. Thesis outline

In this section, we go over the outline of the report. In *Chapter 2*, we delve into ensemble learning and the principle of model selection strategies within such systems. We detail the use case of the Deloitte DocQMiner AI system to highlight the relevance of ensemble learning in information extraction tasks and the importance of considering energy consumption within these systems.

Following, *Chapter 3* presents a comprehensive literature review encompassing the domains of Green AI, model selection strategies, and their intersection with ensemble learning. It synthesizes the existing research and situates our study within the broader context of energy-efficient AI development.

In *Chapter 4*, we outline the strategies for model selection—*Static* and *Dynamic*- and our proposed *Energy-Aware* approach. Additionally, we elaborate on the data sources used, experimental setup, performance evaluation metrics, and the measurement of energy consumption during the inference phase of the models.

In *Chapter 5* we present the results of our experiments, comparing the performance of the Static and Dynamic model selection strategies to the Full Ensemble baseline across the two datasets. The chapter includes a detailed analysis of the precision-recall trade-offs and the resource efficiency gains achieved by each strategy.

[1]The replication package for this study is available at the following DOI link: 10.6084/m9.figshare.25481269

In *Chapter 6*, we discuss the implications of our findings for Green AI and the practical considerations for implementing these strategies in the industry. We also address our study's limitations and suggest avenues for future research.

Concluding the report, in *Chapter 7* we summarize the key findings of the thesis, reaffirming the significance of model selection strategies for achieving a favorable balance between accuracy and energy efficiency in ensemble learning systems. It shows the contribution of this research to the pursuit of sustainable AI practices in production environments.

2

Background

In this section, we highlight the concept of ensemble learning, followed by a detailed study of the use case and the infrastructure of the AI system.

2.1. Ensemble of models

Ensemble learning is a strategic approach that combines several individual and diverse models to achieve better generalization performance [59]. The strength of ensemble learning lies in its diversity; a combination of models can provide a more robust and accurate output than any single model can [15, 24]. Popular applications of ensemble learning are speech recognition [18, 36] and classification [52, 42], image classification [54, 35, 31] and forecasting [44, 38, 7, 49].

A well-known method within ensemble learning is bagging (Bootstrap Aggregating), which enhances the stability and accuracy of machine learning algorithms by training multiple learners on various subsets of the original dataset and then aggregating their predictions to form a final decision [5].

At the core of our research are *model selection strategies* for ensembles. Model selection entails selecting a subset of models from an ensemble of models to optimize for a particular performance metric[8]. We use *model selection* within ensembles to reduce resource consumption in alignment with Green AI principles.

2.2. DocQMiner

To evaluate the impact of using *model selection strategies* for ensemble learning in a live production environment, we use the AI system DocQMiner [16], a proprietary tool owned and developed by Deloitte NL.

This system utilizes a diverse mixture of machine learning (ML) models and NLP technologies to extract, process, and analyze data from textual documents. The tool is an information extraction tool and widely used in the industry across many different domains, for document set sizes ranging from 100 to over 100,000 documents. In this work, we focus on the use case of extracting relevant properties from contracts.

After processing a document, the AI system makes predictions for predefined key properties. Users select properties relevant to their specific use case; for instance, Figure 2.1 illustrates highlighted contract properties like *Title*, *Parties*, *Start Date* and *Goal*. The tool allows users to input a contract and delivers the predictions for the properties of interest. This design makes contract review more efficient, especially for complex documents [41].

Figure 2.1: This figure shows an example of a document with relevant properties highlighted. *Title, Start date, Parties, and Goal* are highlighted to showcase how specific information can be extracted and structured from legal documents.

Users go through a four-step process, as shown in Figure 2.2:

1. **Suggest:** DocQMiner reads the documents and gives suggestions for the properties to extract from the documents.

2. **Analyze and tag:** Using these suggestions, users can quickly go through the documents by either accepting the suggestions or manually selecting the correct text in the document

3. **Review:** Through the four eyes principle done by the experts, the workflow process always contains the correct information.

4. **Learn:** Using the tagged information, DocQMiner learns to provide better suggestions. This further reduces processing time and improves quality.

Figure 2.2: Workflow diagram of DocQMiner: This graphic illustrates the four-step workflow of DocQMiner, starting from unstructured documents to producing structured information. Step 1 involves generating suggestions for the documents, which are then analyzed and tagged in Step 2. The tagged documents undergo a manual review in Step 3, and the system learns from this input to improve future suggestions in Step 4.

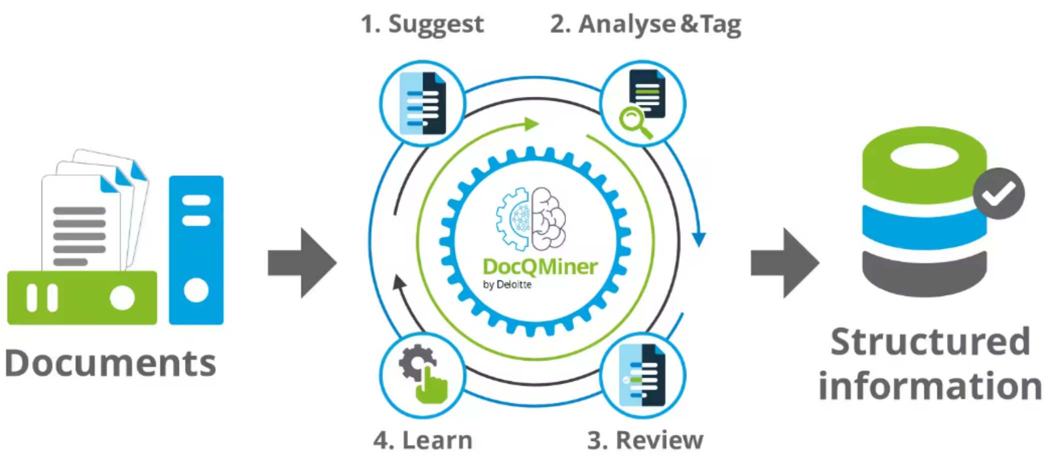
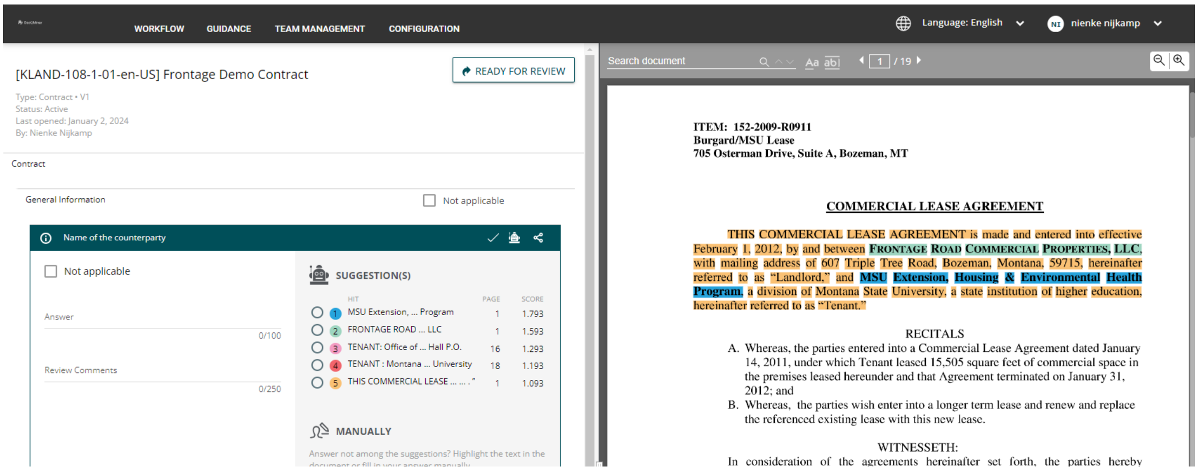


Figure 2.3 shows the user interface of DocQMiner. On the left, the user can see the property and the suggestions, and on the right these suggestions are highlighted in the original document.

Figure 2.3: Interface of DocQMiner: This screenshot showcases the DocQMiner tool in action, highlighting its ability to suggest potential answers for identifying the name of the Counterparty in a commercial lease agreement. The left panel displays the input fields for properties and their suggestions, while the right panel shows the document, with the suggestions highlighted.



The AI system employs a set of diverse (pre-)trained models, as shown in Figure 2.4, to compile a ranked top 5 of textual predictions for every property queried. For every property, each model in the ensemble produces a set of predictions and they are aggregated to form the set of five final predictions. The human-in-the-loop workflow highlighted in Figure 2.2 ensures the validity of the processed data.

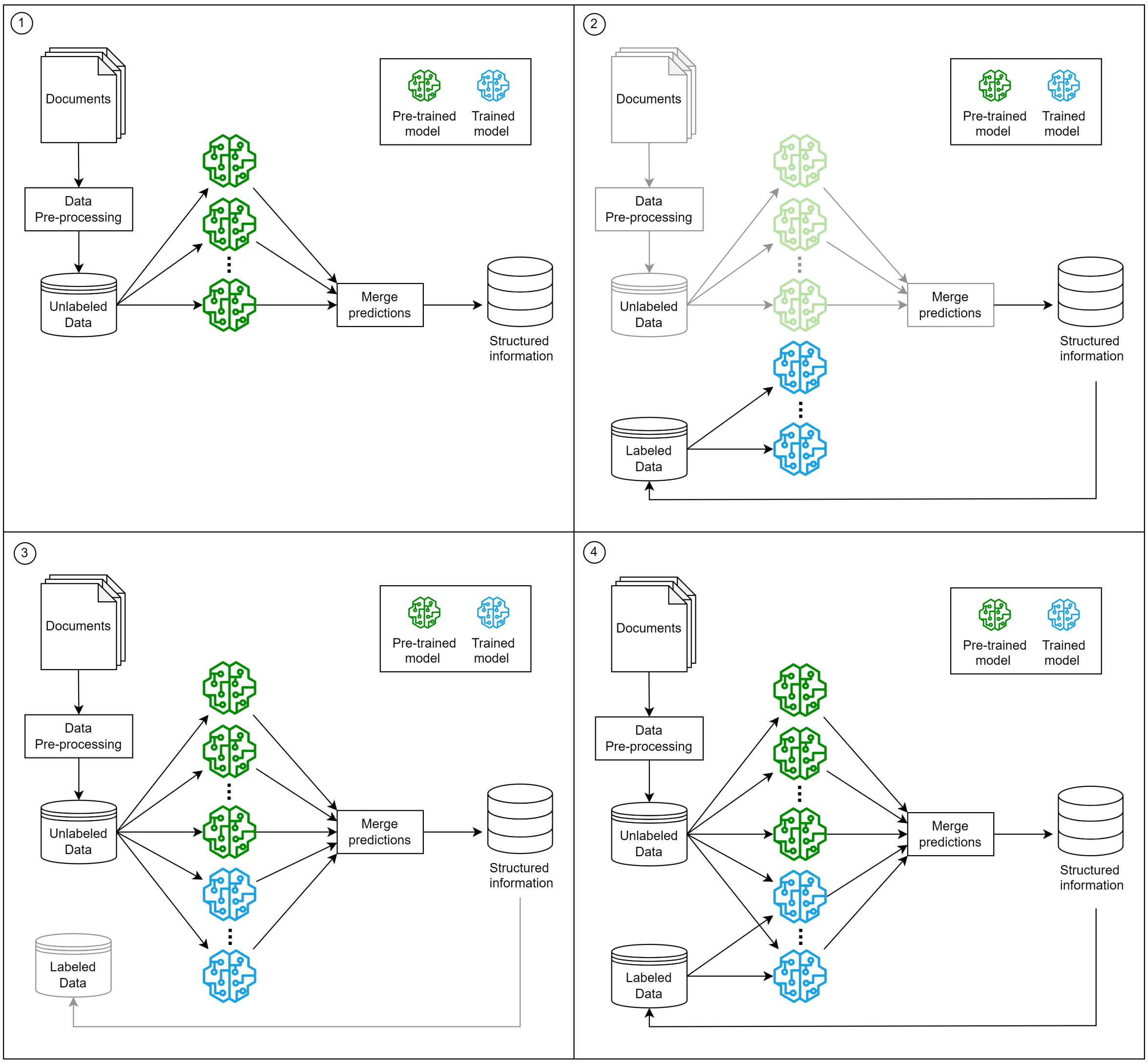


Figure 2.4: Workflow of DocQMiner. (1) Initially, documents are processed using just the pre-trained models. (2) After processing an initial set of documents, models are trained using the processed data. (3) After training models, documents are processed using the pre-trained and trained models. (4) After processing more documents, the trained models can be trained again on the newly processed documents. Steps 3-4 can be continuously repeated.

Figure 2.4 provides a conceptual diagram of the system's processing workflow. Input documents are subjected to a pre-processing step, which ensures that the text within the documents is prepared for subsequent analysis by the system's models.
Out of the box, the ensemble contains a set of pre-trained models to create predictions for the properties. Additionally, once a set of documents within a domain has been completely processed by the system, an additional set of models can be trained using the in-domain data from the processed documents, as depicted by the Labeled Data in Figure 2.4.

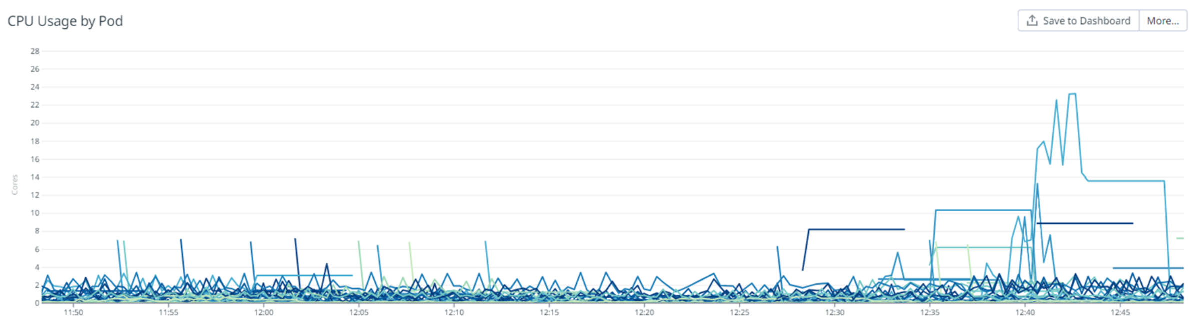
2.3. Energy consumption

This paper analyses the energy consumption of the AI system in a production environment. Therefore, we will leverage the existing monitoring tool within the system, Datadog [14], which offers an extensive set of features for extracting metrics from the system. This cloud monitoring service provides a framework for tracking and analyzing energy use metrics within our AI deployments, as noted in industry literature[40].

Figure 2.5 illustrates a snapshot of metrics (CPU Usage in this case) captured by Datadog at the time of an upload of one document. The peaks in CPU Usage show the impact of operating the DocQMiner model set compared.

DocQMiner uses a substantial amount of CPU when processing a single document. Considering DocQMiner processes document sets ranging in size from 100 to over 100,000 documents per instance, there are environmental and financial implications. This energy usage escalates operational costs and enlarges the carbon footprint of using DocQMiner, challenging the sustainable principles of Deloitte[17].

Figure 2.5: Overview of CPU usage during inference of the model ensemble: This figure displays a snapshot of CPU usage metrics captured by Datadog during the upload of a single document. The peaks in CPU usage highlight the significant computational resources consumed by the DocQMiner model set.



3

Related work

Green AI, the intersection of energy efficiency and model accuracy in Artificial Intelligence (AI) has sparked a growing interest amongst researchers [51, 47]. Our work focuses on reducing resource consumption in ensemble learning using model selection strategies in a live production environment. To the best of our knowledge, there is no other work combining all elements. Below, we highlight the most significant contributions in the areas of model selection strategies and Green AI.

Zhou et al. [60] pivoted model selection in ensemble learning in 2002. Their work introduced GASEN, an approach that begins by allocating initial weights to neural networks, then employs a genetic algorithm to refine these weights. The optimized weights are instrumental in selecting the most effective subset of the ensemble. GASEN proves that incorporating the full ensemble may not always be the optimal strategy, demonstrating the potential of selecting a subset of models.

Li et al. [37] present a novel approach, IRENE, to ensemble learning that focuses on balancing performance and computational cost for inference of ensemble learning. Using a learnable selector, base models, and implementing early halting in a sequential model setup, IRENE reduces inference costs by up to 56% while maintaining comparable performance to full ensembles. IRENE presents a highly effective strategy for ensemble learning in contexts where sequential processing is feasible, as opposed to parallelism. However, the specific use case addressed in this paper necessitates a parallel approach, and as such, does not accommodate the sequential processing model that IRENE requires.

David et al. [15] propose an ensemble learning approach based on the consensus of multiple models for class prediction. It operates under the assumption that once multiple models predict a class it is likely the correct one. This approach significantly reduces computational costs by about 50% while maintaining accuracy. Despite its effectiveness in classification tasks, this approach is not directly applicable to our work, as our ensemble is geared towards information extraction rather than classification, requiring a different methodological framework.

Kotary et al. [32] introduce a framework that combines machine learning and combinatorial optimization for differentiable model selection in ensemble learning. Their method of storing data for predicting optimal subsets in a neural network contrasts with our approach, which involves a more straightforward collection and selection of models within an ensemble for task-specific optimization. Additionally, as with the method from David et al. highlighted above, the success achieved on classification tasks cannot immediately be transferred to information extraction.

Cordeira et al. [12] present a new approach called Post-Selection Dynamic Ensemble Selection (PS-DES). PS-DES evaluates ensembles selected by different DES techniques using different metrics to determine the best ensemble for each query instance. Their work introduces static and dynamic model selection based on evaluating a preliminary set of results. The relevance of these selection methods to our study stems from their design goal of integration into AI pipelines, which aligns with our focus on applying these strategies to a real-life use case. Consequently, we adopt these methods to reduce the number of models during inference in ensembles.

Zhang et al. [58] propose a method named EDDE (Efficient Diversity-Driven Ensemble for Deep Neural Networks) to improve ensemble accuracy while reducing training costs. By selectively transferring generic knowledge and using a diversity-driven loss function, EDDE outperforms other ensemble methods in both Computer Vision and Natural Language Processing tasks. The Boosting-based framework further enhances diversity, making EDDE an efficient and effective ensemble learning approach for neural networks. EDDE proved to reduce the cost of training ensemble models for NLP tasks, however the scope of our research lies within reducing inference cost.

Shazeer et al. [48] discusses the implementation of a Sparsely-Gated Mixture-of-Experts Layer in neural networks, achieving significant improvements in model capacity and computational efficiency. The approach outperforms state-of-the-art models in language modeling and machine translation tasks by addressing challenges such as load balancing and network bandwidth. The hierarchical MoE structure allows for scalability and improved expert utilization. This technique shows promise for enhancing deep learning models with large training datasets. The approaches presented in the research change the training and inference process within the models, while our work proposes an optimization for inference for ensembles without changing the existing models.

Savelka et al. [46] assess GPT-4's performance compared to human annotators in analyzing legal texts. Results show GPT-4 performs similarly to well-trained law student annotators, indicating its capability in this specialized domain. This suggests GPT-4's effectiveness in analyzing legal concepts, offering potential cost and time savings in semantic annotation tasks requiring specialized expertise. The context of extracting text from complex legal documents aligns with the use case of our research, though their work is focused on GPT-4, and our work is focused on the AI system DocQMiner.

Within this landscape of optimizing ensemble learning, the work of Gowda et al. [26] introduces a critical consideration for assessing model efficiency. They propose a metric, the GreenQuotientIndex (GQI), that penalizes high electricity consumption while considering accuracy. The authors conduct a comprehensive study on the electricity consumption of different deep learning models, highlighting the often overlooked trade-off between accuracy and energy efficiency. Our work takes this concept further by demonstrating the practical application of the GQI in real-time production environments that rely on model ensembles.

Wu et al. [57] in collobaration with Facebook explore the environmental impact of AI, focusing on data, algorithms, and system hardware. They highlight the exponential growth in AI and the need for environmentally-responsible advancements. Key points include the carbon footprint of AI computing, challenges in data scaling, and the importance of optimizing AI infrastructure for sustainability. Recommendations include developing resource-efficient models, maximizing accelerator utilization, and incorporating environmental sustainability in AI system design. Overall, the document emphasizes the importance of addressing the environmental implications of AI growth through holistic and sustainable approaches.

Henderson et al. [29] introduces a framework for tracking the energy and carbon footprints of machine learning experiments, aiming to promote energy-efficient research. It highlights the complexities and inaccuracies of current estimation methods, emphasizing the importance of accurate reporting for driving mitigation strategies. The framework enables easy reporting of energy and carbon metrics, facilitating the creation of leaderboards to incentivize energy-efficient algorithms. Immediate actions include utilizing the framework for reporting and participating in energy efficiency leaderboards to promote sustainable machine learning practices. This work highlights the continued need for live monitoring in AI systems, and how monitoring can entice AI practitioners to be more energy conscious when developing AI systems.

Schwartz et al. [47] concept of Green AI, advocating for more environmentally friendly and inclusive AI research. It highlights the increasing computational costs of deep learning models and proposes measuring efficiency using Floating Point Operations (FPO). Researchers can reduce computational expenses by reporting FPO and promoting efficiency as an evaluation criterion without sacrificing performance. The paper emphasizes making AI more sustainable and accessible, encouraging a shift towards Green AI practices.

The review of Green AI literature by Verdecchia et al. [53] shows a growing interest in AI models' energy efficiency, focusing on monitoring, hyperparameter tuning, and model benchmarking. Most studies target the training phase, with significant energy savings reported. Industry involvement is limited, and tools for Green AI are scarce. The results suggest a need to bridge academic research with industrial practice for broader impact.

Research by Salveka et al. [45] discusses using pre-trained language models to identify explanatory sentences in legal case decisions. It highlights the effectiveness of transformer-based models in detecting useful sentences for explaining legal concepts. The models outperform traditional approaches and can learn sophisticated linguistic features. The results suggest potential applications in legal information retrieval and statutory interpretation. This work provides insights for improving legal text analysis and understanding statutory terms. The work uses the same dataset for legal information extraction, the CUAD dataset [30]. However, the performance was reported using the normalized discounted cumulative gain instead of the F1 score in our work.

We have compiled an overview of the relevant literature highlighted in this chapter and their main topics relevant to our work in Table 3.1. The overview shows that the works address some topics; however, the knowledge gap addressed in this research combines all topics.

Table 3.1: This table shows significant research papers across various domains of artificial intelligence, categorized under Efficiency/Green AI, Ensemble Learning, Subset of Models for Inference, Information Extraction, and Industry Involvement. Our work, ”Green AI in Action: Strategic Model Selection for Ensembles in Production,” integrates all these aspects.

Paper			Topic				
Authors	Title	Year	Efficiency / Green AI	Ensemble learning	Subset of models for inference	Information extraction	Industry involvement
Zhou et al. [60]	Ensembling neural networks: Many could be better than all	2002		✓	✓		
Shazeer et al. [48]	Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer	2017	✓	✓	✓		
Zhang et al. [58]	Efficient Diversity-Driven Ensemble for Deep Networks	2020	✓	✓		✓	
David et al. [15]	Adaptive consensus-based ensemble for improved deep learning inference cost	2021	✓	✓	✓		
Savelka et al. [45]	Discovering explanatory sentences in legal case decisions using pre-trained language models	2021				✓	
Wu et al. [57]	Sustainable AI: Environmental implications, challenges and opportunities	2022	✓				✓
Li et al. [37]	Towards Inference Efficient Deep Ensemble Learning	2023	✓	✓	✓		
Kotary et al. [32]	Differentiable model selection for ensemble learning	2023		✓	✓		
Cordeira et al.[12]	A post-selection algorithm for improving dynamic ensemble selection methods	2023		✓	✓		
Gowda et al. [26]	Watt For What: Rethinking Deep Learning’s Energy-Performance Relationship	2023	✓				
Savelka et al. [46]	Can GPT-4 Support Analysis of Textual Data in Tasks Requiring Highly Specialized Domain Expertise?	2023				✓	
Verdecchia et al. [53]	A systematic review of Green AI	2023	✓				
Henderson et al. [29]	Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning	2023	✓				✓
This research	*Green AI in Action: Strategic Model Selection for Ensembles in Production*	*2024*	✓	✓	✓	✓	✓

4

Methodology

This section outlines the methodology and evaluation of our *model selection strategies*. We highlight the two selection strategies, *Static* [39] and *Dynamic* [12], and describe our approach to using them to reduce resource consumption even more efficiently, *Energy-Aware selection*.

We continue with data collection and analysis, followed by the experimental setup and the performance evaluation metrics. Lastly, we explore the energy consumption of the models during inference.

4.1. Model Selection

This study aims to analyze the impact of implementing model selection strategies in a live AI system that uses an ensemble of models on its energy usage and accuracy. In this section, we highlight how two selection strategies, *Static* [39] and *Dynamic* [12], can be used for this, and we describe our approach to using them to reduce resource consumption even more efficiently, *Energy-Aware selection*.

4.1.1. Model selection for energy efficiency

In pursuit of sustainable AI practices, our study assesses existing *Static* [39] and *Dynamic* [12] model selection strategies to reduce energy consumption in a live AI system. These strategies are used to reduce the number of models or improve the efficiency of model usage used during inference, which is a significant determinant of overall energy usage [37].

The *Static* strategy selects an optimal subset of models for general tasks across the domain, while the *Dynamic* strategy adapts model selection to the specifics of each task, aiming to conserve energy without compromising the system's accuracy.

Our main contribution lies in the novel *Energy-Aware selection* approach, which enhances the standard *Static* and *Dynamic* strategies by integrating an energy-aware metric in their application. This metric informs the selection process, ensuring that only the most energy-efficient models are chosen for the task at hand.

4.1.2. Static selection

Static selection involves choosing the optimal subset of models for an entire domain. This approach generalizes the unique characteristics of the domain and selects the optimal subset across all queried properties. The following equation shows the process of *Static* selection.

$$S_{a^*} = argmax_{S_i \in S} F1(S_i) \tag{4.1}$$

where S is the set of all possible model subsets, F1(S_i) is the F1 score of subset S_i on the training set, and S_{a^*} as the optimal model subset chosen for evaluation.

4.1.3. Dynamic selection

Dynamic selection is based on the belief that a model subset might not be optimal for an entire domain, but more specifically for the properties within the domain [12]. Therefore, we take the optimal subset for every queried property within the domain. This approach could be beneficial as the selection is more optimized per property. The following equation shows the process of *Dynamic* selection.

$$S_{a^*}(p_j) = argmax_{S_i \in S} F1(S_i, p_j) \tag{4.2}$$

We denote P as the set of all properties within the domain. $F1(S_i, p_j)$ is the F1 score of subset S_i for property p_j on the training set. $S_{a^*}(p_j)$ is the optimal model subset for property p_j in P chosen for evaluation.

4.1.4. Energy-Aware selection

Both selection strategies, *Static* and *Dynamic*, should reduce energy consumption because a subset of models is used instead of all. However, neither strategy considers how different models compare in terms of energy efficiency. For instance, one model could slightly improve accuracy while costing a significantly larger amount of energy than another.

We propose an enhancement to the *Static* and *Dynamic* approach that discounts accuracy with resource consumption in the selection of models, *Energy-Aware selection*. We use the Green-QuotientIndex (GQI) [26] to factor the trade-off between accuracy and electricity usage. We add GQI to both *Static* and *Dynamic* versions, and compute it as follows:

$$GQI_{static} = \beta \times \frac{F1(S_i)^{\alpha}}{log_{10}(C(S_i))} \tag{4.3}$$

$$GQI_{dynamic} = \beta \times \frac{F1(S_i, p_j)^{\alpha}}{log_{10}(C(S_i))} \tag{4.4}$$

This metric evaluates the trade-off between accuracy and power consumption, where α and β are constants used to scale the GQI. The power consumption ($C(S_i)$) can vary significantly across different models, and therefore the logarithm of the power consumption is taken. Not all accuracy points ($F1(S_i)$ for GQI_{static} and $F1(S_i, p_j)$ for $GQI_{dynamic}$) have the same weight, as it is much easier to get from 0.4 to 0.5 then it is to go from 0.8 to 0.9. Therefore, the power (constant α) of the accuracy reflects the difference in difficulty.

Through the previously mentioned monitoring tool, Datadog [14], there is no availability for power consumption. We, therefore, use CPU usage as a proxy to discount the accuracy. We use this way of discounting the accuracy scores for both the *Dynamic* and the *Static* selection approaches highlighted above.

4.2. Data

The datasets for our experiments are selected based on a set of criteria. Each document within the dataset must contain contracts with a complete text, its associated properties, and annotated responses corresponding to these properties. We employ open-source datasets for our study, enhancing the reproducibility of our results. The CUAD [30] dataset and a dataset from the work of Leivaditi et al. [34] are used in this study, both of which were designed to optimize contract review processes and improve the effectiveness of information extraction.

4.2.1. CUAD

The Contract Understanding Atticus Dataset[30] (CUAD) is an extensive corpus tailored for commercial legal contract analysis. It consists of more than 13,000 labels from 510 contracts divided into 25 contract types. Each document contains one or more of 41 distinct properties. The dataset presents a challenging research benchmark that can be used to enhance deep learning models' performance for contract analysis/understanding [9].

The dataset is designed to automate the process of contract review, traditionally a time-consuming and expensive task when performed manually [30]. It also aims to evaluate the generalization capabilities of NLP models in specialized domains.

This dataset is compiled by a team consisting of law students, practicing lawyers, and machine learning experts. Before beginning their annotation work, all team members underwent a training program lasting between 70 to 100 hours. To ensure accuracy and consistency, at least three other team members reviewed and validated every annotation.

4.2.2. Lease Contracts

Leivaditi et al. [34] introduced a specialized benchmark dataset focused on lease agreement documents. These documents were sourced from a publicly available dataset by the U.S. Securities and Exchange Commission (SEC, 2020). The dataset concentrates on extracting specific properties, including information about the lessor and details of the leased space. This dataset is specifically created to improve the efficiency and accuracy of contract review.

4.2.3. Analysis of datasets

We compare and analyze the datasets employed to understand the characteristics of the datasets and ensure the validity of the results. Table 4.1 shows the comparison of relevant characteristics. The CUAD and Lease Agreement datasets exhibit similar document lengths, with an average of approximately 8,000 words per document, as indicated by the *#WordsPerDocument* metric. Consequently, we will not regard document length as influencing our results.

Table 4.1: Comparison of characteristics between CUAD and Lease Agreement Datasets: This table provides a detailed comparison of the CUAD and Lease Agreement datasets, including the number of documents, types of documents, average words per document, number of annotated properties, total number of queries, number of missing annotations, and the percentage of missing annotations.

Dataset	CUAD [30]	Lease Agreement [34]
#Documents	510	123
#TypesOfDocuments	**25**	**1**
#WordsPerDocument	7861	8053
#AnnotatedProperties	41	12
#TotalQueries	20,910	1476
#MissingAnnotations	13,959	494
MissingAnnotation (%)	**66.77**	**33.47**

The CUAD dataset encompasses a diverse collection of 25 types of contracts, reflecting a wide range of legal agreements. In contrast, the Lease Agreement dataset focuses solely on a single type of contract, specifically lease agreements. With the CUAD dataset's variety, the model selection process must account for the nuances and intricacies inherent in different types of contracts. Conversely, the Lease Agreement dataset's singular focus may allow for more specialized model tuning and optimization tailored specifically to lease agreements.

Additionally, there is a notable difference in the number of annotated properties between the two datasets. With *#MissingAnnotations*, we indicate the number of properties that were queried but did not have a ground truth in the document, thus missing an annotation. Specifically, 66.77% of the queried properties in the CUAD dataset lack annotations, in contrast to the Lease Agreement dataset, which is missing annotations for only 33.47% of its properties. Despite the significant number of missing annotations, we use these datasets to cover a broader array of use cases, including documents where the answer might not always exist. The varied percentage of missing annotations allows us to cover more ground in our results.

4.3. Experimental setup

To integrate models tailored for specific domains, we select a subset of documents from the CUAD [30] and Lease Agreement [34] datasets that reflect the entire dataset. We process, annotate, and train models on these documents. After training, we run an evaluation on a held-out test set.

For both approaches, we run inference using the complete model ensemble on the documents from the validation set. We then evaluate the performance of the entire ensemble and each subset of models within the ensemble. The subset that demonstrates the highest performance, as determined by F1 scores from the training set, is selected for further testing.

This optimally performing subset is applied to the test set documents to evaluate its effectiveness on new data. To ensure the validity of the results, we create the validation and test datasets through 5-fold cross-validation.

As implemented in the tool, the predictions made by the entire model ensemble establish the performance evaluation baseline.

4.4. Performance Evaluation

When paired with the annotations, the ensemble's predictions can result in various scenarios: True Positive, where the prediction aligns with the annotation; False Positive, where a prediction exists but fails to match the annotation; and False Negative, where a prediction exists but no corresponding annotation for the property exists.

To evaluate the performance of the strategies and compare it to the baseline situation of using all the models, we use the F1 score at k. F1 score is a balanced metric of Recall and Precision [43].

Precision: Precision [50] is the percentage of all predictions that match an annotation for the top k predictions:

$$Precision@k = \frac{TP}{TP + FP} \quad \text{for the top k predictions} \tag{4.5}$$

Recall: Recall [50] is the percentage of correctly predicted annotations for the top k predictions:

$$Recall@k = \frac{TP}{TP + FN} \quad \text{for the top k predictions} \tag{4.6}$$

Precision discounts for the number of False Positives and Recall for the number of False Negatives. Increasing Precision typically reduces Recall, and vice versa: increasing Recall decreases Precision. This trade-off underscores the importance of prioritizing one over the other based on specific objectives.
F1 score: F1 score provides a balanced measure of Precision and Recall for the top k predictions. To ensure the robustness and reliability of our research, we select a subset of models based on the F1 score.

$$F1@k = 2 \cdot \frac{Precision@k \cdot Recall@k}{Precision@k + Recall@k} \tag{4.7}$$

DocQMiner[16] only shows the top 5 predictions produced by the set of models. Consequently, we evaluate the results based on k set to five. F1@5 serves as our primary criterion for selecting among different strategies, while we report Precision@5 and Recall@5 to provide a detailed view of the factors contributing to the F1@5 score.

4.5. Resource consumption during inference

To find the most energy-efficient model subset, we need a measurement of the models' consumption during inference. DocQMiner is a live-production environment; therefore, we want to measure the system's resource consumption live. Due to this limitation, we focused on tracking the Central Processing Unit (CPU) usage. With the use of Datadog [14], we can record the CPU usage per second for each process, which allows us to isolate the CPU usage per second, specifically during the inference phase of the models. By taking the cumulative sum of the CPU usage per second over the total duration of the process, we obtain the CPU seconds per process [4]:

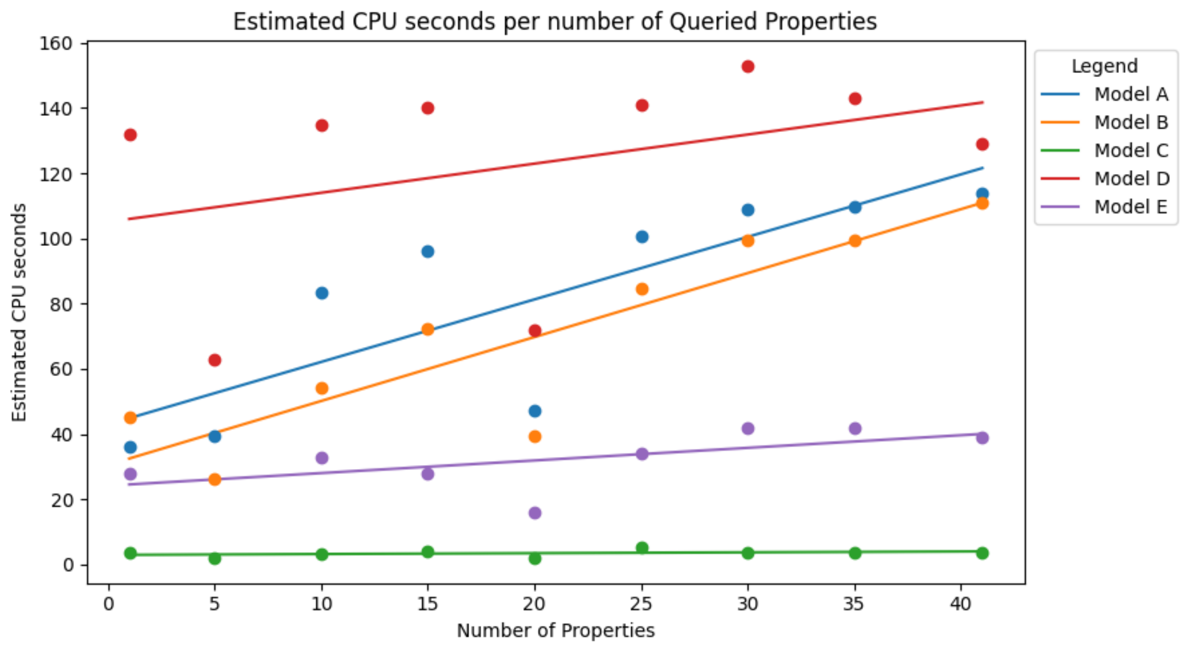
$$\textbf{CPU seconds} = \sum_{i=1}^{n} u_i \tag{4.8}$$

n is the total number of seconds for the process, and u_i is the CPU usage per second during i. Datadog[14] performs this sum calculation.

This setup shows a clear view of each model's performance, given that any other model does not influence the CPU utilization of one model. We gather the data for processing each document and each model for a set of intervals of amount of queried properties, as shown in Figure 4.1.

Given that these measurements are taken in a live production environment, we designed our approach to yield results that closely represent the most probable outcomes. We acknowledge the inherent variability of a production environment, so we plot the measurements' outcomes to account for the variance in results. To ensure the validity of the collected data, we repeat the measurement per number of queried properties 30 times.

Figure 4.1: Estimated CPU seconds per number of Queried Properties: This figure illustrates the relationship between the number of queried properties and the CPU seconds required for each model in a live production environment. The measurements were taken using Datadog to record the CPU usage per second during the inference phase. Each data point represents the average CPU seconds for processing documents with a specific number of queried properties, averaged over 30 repetitions to account for the variability in a production environment.



5

Results

Our evaluation of selection strategies across two datasets—CUAD and Lease Agreement—reveals significant differences in accuracy metrics, including Precision at 5 (P@5), Recall at 5 (R@5), F1 score, and the number of correct and incorrect predictions made by the models involved (see Table 2). *Full Ensemble* reflects the baseline of using all the models for every property within all documents.

5.1. Experimental context

5.1.1. Legal information extraction

Extracting relevant information from legal documents presents significant challenges due to their complex nature. These texts often require identifying specific details within extensive documents. This task notably differs from more straightforward tasks like classification, where an F1 score below 0.6 might be deemed insignificant. In the context of legal text analysis, the F1 scores are typically lower, reflecting the intricate nature of the work involved.

These lower scores are supported in research conducted by Savelka et al.[46] using GPT-4 for property extraction from complex legal documents, which reported a Precision of 0.63, a Recall of 0.46, and an F1 score of 0.53. Despite being from a different dataset, these results show that relevant information extraction from legal documents is not a trivial task, and the results from the *Full Ensemble* should be interpreted as such.

5.1.2. Recall-oriented system

DocQMiner [16] is developed prioritizing Recall as the key metric, which aligns with its human-in-the-loop design. This design allows users to choose the best answer from the predictions provided. Consequently, the model development concentrates on accurately identifying rele-

Table 5.1: Comparison of Full Ensemble and Static and Dynamic Strategies for CUAD dataset [30]: This table presents the performance of different strategies in terms of Precision (P@5), Recall (R@5), F1-Score (F1@5), the number of correct and incorrect predictions, and relative CPU usage. The Full Ensemble serves as a baseline, while Static and Dynamic strategies are evaluated in their original forms and with cost inclusion.

Strategy		P@5	R@5	F1@5	# Correct predictions	# Incorrect predictions	Relative CPU Usage
Full Ensemble		*0.0871*	*0.5338*	*0.1495*	*490*	*5928*	*100%*
Static	Original	0.6457	0.4872	0.5544	408	265	60.39%
	Energy-Aware	0.6370	0.4652	0.5366	394	259	26.52%
Dynamic	Original	0.6118	0.5448	0.5750	446	322	71.28%
	Energy-Aware	0.6099	0.5471	0.5756	446	326	67.11%

Table 5.2: Comparison of Full Ensemble and Static and Dynamic Strategies for Lease Agreement dataset[34]: This table presents the performance of different strategies in terms of Precision (P@5), Recall (R@5), F1-Score (F1@5), the number of correct and incorrect predictions, and relative CPU usage. The Full Ensemble serves as a baseline, while Static and Dynamic strategies are evaluated in their original forms and with cost inclusion.

Strategy		P@5	R@5	F1@5	# Correct predictions	# Incorrect predictions	Relative CPU Usage
Full Ensemble		*0.1386*	*0.4448*	*0.2203*	*90*	*584*	*100%*
Static	Original	0.1763	0.3677	0.2545	77	376	64.62%
	Energy-Aware	0.1817	0.2404	0.2054	57	278	0,82%
Dynamic	Original	0.1852	0.4732	0.2652	93	427	81.22%
	Energy-Aware	0.2106	0.3415	0.2588	70	282	47.43%

vant properties rather than minimizing the prediction of incorrect properties.

5.2. Full Ensemble

For CUAD, the *Full Ensemble* strategy achieved a P@5 of 0.0871, R@5 of 0.5338, and F1 of 0.1495, making 490 correct predictions and 5928 incorrect, with 100% CPU usage. For Lease Agreement, *Full Ensemble* had a P@5 of 0.1386, R@5 of 0.4448, and F1 of 0.2203, with 90 correct and 584 incorrect predictions, at 100% CPU usage.

5.3. Static Strategy

For CUAD, the *Static Original* variant, as defined in Section 4.1.2, achieved a P@5 of 0.6457, R@5 of 0.4872, and F1 of 0.5544, with 408 correct and 265 incorrect predictions, consuming 60.39% CPU compared to the *Full Ensemble*. For Lease Agreement, *Static Original* posted a P@5 of 0.1763, R@5 of 0.3677, and F1 of 0.2545, with 77 correct and 376 incorrect predictions, at 64.62% CPU usage.

The *Static Energy-Aware* variation, as defined in Section 4.1.4, for CUAD showed a slight decrease in performance to an F1 of 0.5366, P@5 of 0.6370, R@5 of 0.4652, with 394 correct and 259 incorrect predictions, and reduced energy consumption to 26.52% CPU. For the Lease Agreement dataset, it managed an F1 of 0.2054, P@5 of 0.1817, and R@5 of 0.2404, with 57 correct and 278 incorrect predictions, drastically cutting CPU use to 0.82%.

Compared to the *Full Ensemble* baseline, overall Recall has slightly declined; however, overall Precision has increased, especially for the CUAD dataset. These results show that the *Static* strategy can correctly identify many properties whilst reducing the 'noise' of incorrect predictions.

5.4. Dynamic Strategy

For CUAD, the *Dynamic Original* strategy, as defined in Section 4.1.3, resulted in an F1 score of 0.5750, a P@5 of 0.6118, an R@5 of 0.5448, 446 correct and 322 incorrect predictions, at 71.28% CPU consumption. For Lease Agreement, it achieved an F1 score of 0.2652, a P@5 of 0.1852, an R@5 of 0.4732, 93 correct and 427 incorrect predictions, with 81.22% CPU usage compared to the *Full Ensemble*.

The *Dynamic Energy-Aware*, defined in Section 4.1.4, showed for CUAD an F1 of 0.5756, P@5 of 0.6099, and R@5 of 0.5471, with 446 correct and 326 incorrect predictions, lowering CPU usage to 67.11%. For the Lease Agreement dataset, the F1 was 0.2588, P@5 of 0.2106, R@5 of 0.3415, with 70 correct and 282 incorrect predictions, reducing CPU consumption to 47.43%.

Overall, the *Dynamic Original* strategy outperforms the baseline on both Precision and Recall. The increase in Recall is notable, considering DocQMiner is tailored to Recall. The increase in Recall is suspected to be due to how the ensemble 'dilutes' the predictions, and the *Dynamic* strategy specializes in specific properties.

5.5. Strategy selection

Identifying an optimal strategy for an ensemble of models hinges on a few considerations. The evaluation of the *Full Ensemble*, *Static*, and *Dynamic* strategies across the CUAD and Lease Agreement datasets provides valuable insights into their respective strengths and weaknesses.

5.5.1. Precision - Recall trade-off

For Precision The *Static Original* strategy stands out in the CUAD dataset with a Precision (P@5) of 0.6457 and an F1 score of 0.5544, significantly reducing the noise of incorrect predictions. Similarly, the Lease Agreement dataset performs with a Precision of 0.1763, compared to 0.1386 Precision for the *Full Ensemble*. These numbers suggest that the *Static Original* strategy offers a solution when minimising false positives is the goal.

For Recall The *Dynamic Original* strategy shines by delivering a Recall (R@5) of 0.5448 for CUAD and 0.4732 for Lease Agreement, coupled with the highest F1 scores (0.5750 and 0.2652). This strategy ensures that more relevant properties are captured, making it ideal when maximising the number of accurately identified properties, which is the goal.

5.5.2. Resource efficiency

In a resource constraint environment, their performance and efficiency balance should inform the choice between the *Static Energy-Aware* and *Dynamic Energy-Aware* strategies.

Extreme Efficiency The *Static Energy-Aware* strategy is unparalleled, especially evident in the Lease Agreement dataset, with CPU usage reduced to nearly 1%. This strategy is suitable for projects where every bit of computational resource saved makes an impact, even at the expense of some accuracy.

Balanced Approach The *Dynamic Energy-Aware* strategy, while not as resource-efficient as its *Static* counterpart, offers a balance between accuracy and resource usage. This balance makes it ideal for scenarios where a moderate resource reduction is acceptable if it means retaining a higher level of accuracy.

5.5.3. Specificity of properties

If the task involves identifying particular properties within legal documents, the specialisation afforded by the *Dynamic* strategies might yield better results. The *Dynamic* strategies are fine-tuned to identify specific properties more effectively, possibly at the cost of broader applicability that *Static* strategies can offer.

However, the success of the *Dynamic* approach hinges on the availability of sufficient information about the specific properties within the test set to identify the optimal subset accurately. In cases where such specific information is unavailable, the *Static* strategy may be the more suitable option, balancing the need for broader coverage with the available data.

5.6. Impact of selection strategies

Our results indicate that implementing model selection strategies can significantly impact both the accuracy and resource consumption of ensemble learning systems. The evaluation of the *Static* strategy suggests a notable improvement in Precision, reducing the number of incorrect predictions. In comparison, the *Dynamic* strategy excels in Recall, effectively retrieving more relevant instances while achieving a reduction in energy consumption, making it ideal for applications where capturing as much relevant information as possible is critical.

Furthermore, by enhancing these strategies with the cost-inclusive GreenQuotientIndex (GQI) [26], we have demonstrated a method for reducing the models' energy consumption without substantially sacrificing accuracy. For instance, the *Static Energy-Aware* strategy decreases the average energy usage from approximately 62% to 14% compared to the *Full Ensemble*, highlighting the potential for significant energy savings. In the case of the *Dynamic Energy-Aware* strategy, we observed an average reduction from around 76% to 57%, showing the effectiveness of this approach in balancing performance with energy efficiency.

These findings confirm that model selection strategies, in their original form and particularly our approach augmenting the original form with an energy consumption metric, can combine the objectives of maintaining high accuracy while reducing the energy demands of AI systems in a production environment.

6 Discussion

This study presents several insights for strategic model selection for ensemble learning in live AI systems, particularly in performance optimization and computational efficiency.

6.1. (Green) AI

For AI practitioners, the findings emphasize the importance and viability of balancing accuracy with computational cost. In the current context, where computational efficiency is both an economic and environmental concern, the study's insights suggest that achieving this balance—maintaining high levels of accuracy while being mindful of the computational resources consumed—is essential for sustainable AI development [47].
These insights offer tangible benefits from an industry standpoint, particularly in strategizing AI developments. The potential for resource-efficient model selection without significantly compromising performance paves the way for greener AI solutions. Such solutions are particularly valuable in resource-intensive applications, contributing to efforts to reduce AI technologies' environmental footprint.

6.2. Future of ensembles

With the rising popularity of large language models (LLMs), one might wonder whether "old-school" ensembles are still the way to go. Research has shown that for most NLP tasks considered state-of-the-art fine-tuned models like TULRv6 generally outperform LLMs by a considerable margin, especially in languages other than English[1]. An LLM may not provide superior solutions for the specific task of legal information extraction.

In addition to accuracy, we prioritize efficiency as a crucial metric. A straightforward auto-regressive model like BERT_{Large}, which consists of approximately 340 million parameters[20], requires four days of training on 16 TPU processors[19]. In contrast, GPT-4 is rumored to have 1.7 trillion parameters and demands 90-100 days of training on 25,000 GPU processors[21]. Although OpenAI has not officially disclosed the energy consumption of these models, it is reasonable to assume that their resource usage is significant. From the perspective of energy consumption, ensembles of fine-tuned models are preferable to large language models.

6.3. Monitoring

Monitoring is critical to ensuring the sustainability of AI systems [29, 57]. Practitioners need to be able to monitor their applications to understand their energy consumption and environmental impact. Tools and interfaces that facilitate the development of AI systems should also incorporate features for monitoring energy usage. Monitoring tools would enable practitioners to create eco-friendly systems without significant effort, addressing the gap in making sustainable AI more accessible and practical.

Moreover, by identifying the major energy consumers within live AI systems, practitioners can focus on reducing energy consumption in the most impactful areas, leading to more efficient and sustainable AI solutions.

6.4. Practical implications

Balancing theoretical performance with practical considerations is vital when choosing the best model for a task. The *Static* strategy is typically more straightforward to deploy and compatible with a broader range of infrastructures, making it a practical option for many setups. The *Dynamic* strategy, while potentially more effective for specific tasks, might demand a more complex infrastructure setup. Therefore, the decision should consider the desired accuracy, efficiency, and practicality of integrating and maintaining the strategy into existing systems.

Companies should prioritize sustainable AI development as the impact of AI on the workforce continues to grow, accompanied by significant financial and environmental costs. Although the Green AI field is expanding, its connection to the industry remains insufficient [53]. This research demonstrates that making AI systems in production more sustainable is feasible. Continued efforts in this direction can further convince companies of the practicality and benefits of adopting sustainable AI practices.

6.5. Limitations

While this study provides valuable insights, it also has limitations. The evaluation was conducted on two specific datasets within the legal domain, which may limit the generalizability of the findings to other types of documents or domains. Future research could explore the applicability of these model selection strategies across a broader range of datasets and domains, not only within information extraction but also in other areas beyond the scope of NLP.

Additionally, our approach to discounting model selection based on resource consumption relies on CPU seconds as a proxy for energy consumption. Future studies could incorporate more direct metrics of energy consumption, such as power usage in kWh, to assess the environmental impact and reduction in the carbon footprint of different model selection strategies. This limitation aligns with the continued need for accurate and easy-to-use monitoring tools for AI systems [29, 57] highlighted in Section 6.3.

Lastly, our study does not account for the energy costs associated with training or fine-tuning models for specific domains within the ensemble. Future research should investigate whether the accuracy improvements from domain-specific training, see blue models in Figure 2.4, justify these increased energy expenditures. This analysis could provide deeper insights into the efficiency and effectiveness of employing specialized models within ensemble systems.

7 Conclusion

In response to the escalating energy demands of Artificial Intelligence (AI) systems, particularly those employing ensemble learning [37], our empirical study explores model selection strategies to optimize accuracy and energy efficiency. Our research introduces and evaluates two model selection strategies, *Static* and *Dynamic*, aimed at optimizing the performance of ensemble learning systems while minimizing their energy usage.

By evaluating the *Static* and *Dynamic* strategies across the CUAD and Lease Agreement datasets, we have highlighted the adaptability and potential of these approaches to meet diverse needs. Our results reveal that the *Static* strategy improves the F1 score beyond the baseline, reducing average energy usage from 100% from the full ensemble to 62%. The *Dynamic* strategy further enhances F1 scores, while using on average 76% compared to 100% of the full ensemble.

Additionally, we propose an approach that discounts accuracy with resource consumption, the *Energy-Aware* approach, showing potential for further reducing energy usage without significantly impacting accuracy. This method further decreased the average energy usage of the *Static* strategy from approximately 62% to about 14%, and for the optimal *Dynamic* strategy, from around 76% to roughly 57%.

Our findings, especially the successful application of the *Energy-Aware* approach, align with the principles of Green AI [47, 51], advocating for sustainable AI practices that maintain high-performance standards.

This field study on DocQMiner, an AI system actively used in the industry, highlights our research's real-world applicability and significance in advancing sustainable, efficient AI technologies for live production environments.

These insights provide a valuable perspective for the industry on developing AI in a resource-conscious yet effective manner. They highlight the feasibility of using model selection strategies to balance accuracy and computational efficiency, demonstrating the crucial need for adopting strategies that account for accuracy and environmental impact for ensuring sustainable development as AI progresses.

References

[1] Kabir Ahuja et al. “Mega: Multilingual evaluation of generative ai”. In: *arXiv preprint arXiv:2303.12528* (2023).

[2] Lasse F Wolff Anthony, Benjamin Kanding, and Raghavendra Selvan. “Carbontracker: Tracking and predicting the carbon footprint of training deep learning models”. In: *arXiv preprint arXiv:2007.03051* (2020).

[3] Robert M Bell and Yehuda Koren. “Lessons from the Netflix prize challenge”. In: *Acm Sigkdd Explorations Newsletter* 9.2 (2007), pp. 75–79.

[4] James F Brady. “Virtualization and cpu wait times in a linux guest environment”. In: *J. Comp. Resource Mgmt* 116 (2005), pp. 3–8.

[5] Leo Breiman. “Bagging predictors”. In: *Machine learning* 24 (1996), pp. 123–140.

[6] Leo Breiman. “Random forests”. In: *Machine learning* 45 (2001), pp. 5–32.

[7] Salvatore Carta et al. “A multi-layer and multi-ensemble stock trader using deep learning and deep reinforcement learning”. In: *Applied Intelligence* 51 (2021), pp. 889–905.

[8] Rich Caruana et al. “Ensemble selection from libraries of models”. In: *Proceedings of the twenty-first international conference on Machine learning*. 2004, p. 18.

[9] Yupeng Chang et al. “A survey on evaluation of large language models”. In: *arXiv preprint arXiv:2307.03109* (2023).

[10] Tianqi Chen and Carlos Guestrin. “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.

[11] Andrew A Chien et al. “Reducing the Carbon Impact of Generative AI Inference (today and in 2035)”. In: *Proceedings of the 2nd Workshop on Sustainable Computer Systems*. 2023, pp. 1–7.

[12] Paulo RG Cordeiro, George DC Cavalcanti, and Rafael MO Cruz. “A post-selection algorithm for improving dynamic ensemble selection methods”. In: *arXiv preprint arXiv:2309.14307* (2023).

[13] Kate Crawford. *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press, 2021.

[14] DataDog. *DataDog: Cloud Monitoring as a Service*. `https://www.datadoghq.com/`. Accessed: 2024-02-08. 2024.

[15] Nelly David and Nathan S Netanyahu. "Adaptive consensus-based ensemble for improved deep learning inference cost". In: *International Conference on Artificial Neural Networks*. Springer. 2021, pp. 330–339.

[16] Deloitte. *DocQMiner - Deloitte's Solution for Document Processing and Information Extraction*. `https://www2.deloitte.com/nl/nl/pages/risk/solutions/docqminer.html`. Accessed: 2024-02-11. 2022.

[17] Deloitte Netherlands. *Environmental Impacts - Deloitte Netherlands Annual Report 2022/2023*. `https://annualreport.deloitte.nl/annual-report-20222023/nonfinancial-statements--/2-environmental-impacts`. Accessed: 2024-03-27. 2023.

[18] Li Deng et al. "Use of kernel deep convex networks and end-to-end learning for spoken language understanding". In: *2012 IEEE Spoken Language Technology Workshop (SLT)*. IEEE. 2012, pp. 210–215.

[19] Hugging Face. *BERT 101*. `https://huggingface.co/blog/bert-101`. Accessed: 2024-03-28. 2021.

[20] Hugging Face. *Transformers: Pretrained Models*. `https://huggingface.co/transformers/v2.4.0/pretrained_models.html`. Accessed: 2024-03-28. 2020.

[21] Danielle Franca. *GPT4: All Details Leaked*. `https://medium.com/@daniellefranca96/gpt4-all-details-leaked-48fa20f9a4a`. Accessed: 2024-03-28. 2023.

[22] Yoav Freund and Robert E Schapire. "A decision-theoretic generalization of on-line learning and an application to boosting". In: *Journal of computer and system sciences* 55.1 (1997), pp. 119–139.

[23] Jerome H Friedman. "Greedy function approximation: a gradient boosting machine". In: *Annals of statistics* (2001), pp. 1189–1232.

[24] Mudasir A Ganaie et al. "Ensemble deep learning: A review". In: *Engineering Applications of Artificial Intelligence* 115 (2022), p. 105151.

[25] Stefanos Georgiou et al. "Green ai: Do deep learning frameworks have different costs?" In: *Proceedings of the 44th International Conference on Software Engineering*. 2022, pp. 1082–1094.

[26] Shreyank N Gowda et al. "Watt For What: Rethinking Deep Learning's Energy-Performance Relationship". In: *arXiv preprint arXiv:2310.06522* (2023).

[27] Song Han et al. "EIE: Efficient Inference Engine on Compressed Deep Neural Network". In: *2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA)*. 2016, pp. 243–254. DOI: `10.1109/ISCA.2016.30`.

[28] Lars Kai Hansen and Peter Salamon. "Neural network ensembles". In: *IEEE transactions on pattern analysis and machine intelligence* 12.10 (1990), pp. 993–1001.

[29] Peter Henderson et al. “Towards the systematic reporting of the energy and carbon footprints of machine learning”. In: *The Journal of Machine Learning Research* 21.1 (2020), pp. 10039–10081.

[30] Dan Hendrycks et al. “Cuad: An expert-annotated nlp dataset for legal contract review”. In: *arXiv preprint arXiv:2103.06268* (2021).

[31] Gao Huang et al. “Snapshot ensembles: Train 1, get m for free”. In: *arXiv preprint arXiv:1704.00109* (2017).

[32] James Kotary, Vincenzo Di Vito, and Ferdinando Fioretto. “Differentiable model selection for ensemble learning”. In: *International Joint Conference on Artificial Intelligence (IJCAI)*. 2023.

[33] Alexandre Lacoste et al. “Quantifying the carbon emissions of machine learning”. In: *arXiv preprint arXiv:1910.09700* (2019).

[34] Spyretta Leivaditi, Julien Rossi, and Evangelos Kanoulas. “A benchmark for lease contract review”. In: *arXiv preprint arXiv:2010.10386* (2020).

[35] Jichang Li et al. “Semi-supervised deep coupled ensemble learning with classification landmark exploration”. In: *IEEE Transactions on Image Processing* 29 (2019), pp. 538–550.

[36] Sheng Li et al. “Semi-supervised ensemble DNN acoustic model training”. In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2017, pp. 5270–5274.

[37] Ziyue Li et al. “Towards Inference Efficient Deep Ensemble Learning”. In: *arXiv preprint arXiv:2301.12378* (2023).

[38] Fan Liu, Feng Xu, and Sai Yang. “A flood forecasting model based on deep learning algorithm via integrating stacked autoencoders with BP neural network”. In: *2017 IEEE third International conference on multimedia big data (BigMM)*. Ieee. 2017, pp. 58–61.

[39] Gang Luo. “A review of automatic selection methods for machine learning algorithms and hyper-parameter values”. In: *Network Modeling Analysis in Health Informatics and Bioinformatics* 5 (2016), pp. 1–16.

[40] Adam Mckaig and Tahia Khan. “How the Metrics Backend Works at Datadog”. In: (2022).

[41] Afra Nawar et al. “An Open Source Contractual Language Understanding Application Using Machine Learning”. In: *Proceedings of the First Workshop on Language Technology and Resources for a Fair, Inclusive, and Safe Society within the 13th Language Resources and Evaluation Conference*. 2022, pp. 42–50.

[42] Hamid Palangi, Li Deng, and Rabab K Ward. “Recurrent deep-stacking networks for sequence classification”. In: *2014 IEEE China Summit & International Conference on Signal and Information Processing (ChinaSIP)*. IEEE. 2014, pp. 510–514.

[43] David MW Powers. "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation". In: *arXiv preprint arXiv:2010.16061* (2020).

[44] Xueheng Qiu et al. "Ensemble deep learning for regression and time series forecasting". In: *2014 IEEE symposium on computational intelligence in ensemble learning (CIEL)*. IEEE. 2014, pp. 1–6.

[45] Jaromir Savelka and Kevin D Ashley. "Discovering explanatory sentences in legal case decisions using pre-trained language models". In: *arXiv preprint arXiv:2112.07165* (2021).

[46] Jaromir Savelka et al. "Can GPT-4 Support Analysis of Textual Data in Tasks Requiring Highly Specialized Domain Expertise?" In: *arXiv preprint arXiv:2306.13906* (2023).

[47] Roy Schwartz et al. "Green ai". In: *Communications of the ACM* 63.12 (2020), pp. 54–63.

[48] Noam Shazeer et al. "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer". In: *arXiv preprint arXiv:1701.06538* (2017).

[49] Pardeep Singla, Manoj Duhan, and Sumit Saroha. "An ensemble method to forecast 24-h ahead solar irradiance using wavelet decomposition and BiLSTM deep learning network". In: *Earth Science Informatics* 15.1 (2022), pp. 291–306.

[50] Marina Sokolova and Guy Lapalme. "A systematic analysis of performance measures for classification tasks". In: *Information processing & management* 45.4 (2009), pp. 427–437.

[51] Emma Strubell, Ananya Ganesh, and Andrew McCallum. "Energy and policy considerations for deep learning in NLP". In: *arXiv preprint arXiv:1906.02243* (2019).

[52] Gokhan Tur et al. "Towards deeper understanding: Deep convex networks for semantic utterance classification". In: *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 2012, pp. 5045–5048.

[53] Roberto Verdecchia, June Sallou, and Luís Cruz. "A systematic review of Green AI". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* (2023), e1507.

[54] Bin Wang, Bing Xue, and Mengjie Zhang. "Particle swarm optimisation for evolving deep neural networks for image classification by evolving and stacking transferable blocks". In: *2020 IEEE Congress on Evolutionary Computation (CEC)*. IEEE. 2020, pp. 1–8.

[55] Tim Whitaker and Darrell Whitley. "Prune and tune ensembles: low-cost ensemble learning with sparse independent subnetworks". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 8. 2022, pp. 8638–8646.

[56] David H Wolpert. "Stacked generalization". In: *Neural networks* 5.2 (1992), pp. 241–259.

[57] Carole-Jean Wu et al. "Sustainable ai: Environmental implications, challenges and opportunities". In: *Proceedings of Machine Learning and Systems* 4 (2022), pp. 795–813.

[58] Wentao Zhang et al. "Efficient diversity-driven ensemble for deep neural networks". In: *2020 IEEE 36th International Conference on Data Engineering (ICDE)*. IEEE. 2020, pp. 73–84.

[59] Zhi-Hua Zhou. *Ensemble methods: foundations and algorithms*. CRC press, 2012.

[60] Zhi-Hua Zhou, Jianxin Wu, and Wei Tang. "Ensembling neural networks: many could be better than all". In: *Artificial intelligence* 137.1-2 (2002), pp. 239–263.