Industrial end-of-studies project

Presented for the obtainment of the title:

State engineer of Arts and Crafts

Industrial Engineering: Artificial Intelligence & Data Science

---

AI-based tool to support the decarbonization of real estate parcs

---

Prepared by :
Saddik Imad

Presented on June 25, 2024, to a jury consisting of:

Mr. FASSI FIHRI Abdelkader	ENSAM	- Examiner
Mr. HAJJI Tarik	ENSAM	- Reporter
Mr. MASROUR Tawfik	ENSAM	- Academic supervisor
Mr. JOUANE Youssef	CESI LINEACT	- Supervisor
Mr. ABOUELAZIZ Ilyass	CESI LINEACT	- Co-supervisor

Academic year : 2023/2024

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my advisor, Youssef JOUANE, for his invaluable guidance and support throughout my internship. Youssef's expertise in the PV domain has been instrumental in my understanding of the project and its technical complexities. His collaborative approach, treating me more like a colleague than a student, fostered a positive and productive learning environment. I am particularly grateful for his support and patience, never pressuring me to deliver results but always present when I needed his guidance.

I am also deeply grateful to my academic supervisor, Tawfik MASROUR, for his continuous encouragement and insightful feedback throughout the internship. Tawfik's guidance and support were instrumental in shaping my research direction and ensuring the successful completion of the project.

I am also deeply grateful to the CCCA-BTP for financing my internship. Their commitment to innovative projects in the energy domain is admirable, and I am honored to have contributed to this important research.

Furthermore, I would like to express my heartfelt appreciation to my parents for their support throughout my life. Their belief in my abilities and their sacrifices to allow me to come to France for this internship have been invaluable.

I extend my sincere thanks to "es Compagnons du Devoir for providing access to real energy production data in Strasbourg. Their contribution has been crucial in grounding our research in real-world applications.

I am also grateful to the Campus BTP CFA OCCITANIE de Toulouse-Muret for sharing extensive information about their building in Toulouse. Their willingness to collaborate has significantly enhanced the depth and applicability of our research.

I want to thank the Solcast team for giving us free access to meteorological data for researchers. Their dedication to supporting scientific research is greatly appreciated and has been crucial in advancing this project.

Finally, I would like to extend my gratitude to everyone who has contributed, directly or indirectly, to the successful completion of this internship. Your support and encouragement have been instrumental in my personal and professional growth.

Abstract

This internship report explores the development of an Artificial Intelligence (AI) tool for accurately predicting the energy production of Photovoltaic (PV) systems integrated into buildings. Building upon previous research, we significantly expanded the dataset used for model training by generating synthetic data with BimSolar, encompassing a wider range of building types, solar panel models, and geographical locations. This resulted in a dataset ten times larger than the original, improving the model's accuracy and generalizability. Furthermore, we transitioned from a forecasting approach to a prediction problem, allowing for direct estimation of energy production without requiring historical data. This shift enhances the model's flexibility and adaptability to new scenarios. We also integrated the CodeCarbon library to measure the carbon emissions associated with model training, emphasizing the importance of energy-efficient solutions. We evaluated the performance of different machine learning models, finding XGBoost and Decision Tree models to be the most efficient and accurate. Finally, we validated the models on real-world study cases, demonstrating their ability to accurately predict energy production for buildings with diverse PV installations. This work paves the way for more efficient, sustainable, and user-friendly AI-powered tools for predicting PV energy production, ultimately supporting the transition towards a cleaner and more efficient energy future.

Keywords : Renewable energy, Building-integrated photovoltaic, Machine Learning, Energy prediction, Decarbonization.

Résumé

Ce rapport de stage explore le développement d'un outil d'Intelligence Artificielle (IA) pour prédire avec précision la production d'énergie des systèmes photovoltaïque intégrés aux bâtiments. S'appuyant sur des recherches antérieures, nous avons considérablement élargi l'ensemble de données utilisé pour l'entraînement du modèle en générant des données synthétiques avec BimSolar, englobant une plus grande variété de types de bâtiments, de modèles de panneaux solaires et de localisations géographiques. Cela a abouti à un ensemble de données dix fois plus important que l'original, améliorant ainsi la précision et la généralisabilité du modèle. De plus, nous sommes passés d'une approche de prévision à un problème de prédiction, permettant une estimation directe de la production d'énergie sans nécessiter de données historiques. Ce changement améliore la flexibilité et l'adaptabilité du modèle à de nouveaux scénarios. Nous avons également intégré la bibliothèque CodeCarbon pour mesurer les émissions de carbone associées à l'entraînement du modèle, soulignant l'importance de solutions écoénergétiques. Nous avons évalué la performance de différents modèles d'apprentissage automatique, constatant que les modèles XGBoost et Decision Tree étaient les plus efficaces et précis. Enfin, nous avons validé les modèles sur des cas d'étude réels, démontrant leur capacité à prédire avec précision la production d'énergie pour des bâtiments avec diverses installations PV. Ce travail ouvre la voie à des outils alimentés par l'IA plus efficaces, durables et conviviaux pour prédire la production d'énergie PV, soutenant ainsi la transition vers un avenir énergétique plus propre et plus efficient.

Mots-clés : Énergie renouvelable, Photovoltaïque intégré au bâtiment, Apprentissage automatique, Prédiction énergétique, Décarbonisation.

List of Figures

List of Tables

List of Abbreviations

AC Alternating current

AI Artificial Intelligence

IA Intelligence Artificielle

AOCDTF Association ouvrière des Compagnons du Devoir et du Tour de France

AR Augmented Reality

ANN Artificial Neural Networks

BIM Building Information Modeling

BIPV Building-Integrated Photovoltaics

BR Bayesian Regularization

CAVE Cave Automatic Virtual Environment

CAMS Copernicus Atmosphere Monitoring Service

CFA Centre de formation d'apprentis

CNN Convolutional Neural Network

CO2 Carbon dioxide

CCCA-BTP Comité de concertation et de coordination de l'apprentissage du bâtiment et des travaux publics

DC Direct current

DL Deep Learning

DNN Dense neural network

DNI Direct Normal Irradiance

DHI Diffuse Horizontal Irradiance

EU European Union

FFNN Feedforward neural network

GBR Gradient Boosting Regressor

GHI Global Horizontal Irradiance

GR Gaussian regression

GRNN Generalized Regression Neural Network

GRU Gated Recurrent Unit

IFC Industry Foundation Classes

IEA International Energy Agency

IIoT Industrial Internet of Things

JENII Jumeaux d'Enseignement Numériques, Immersifs et Interactifs

LM Levenberg-Marquardt

LSTM Long Short-Term Memory

MAE Mean Absolute Error

MAPE Mean Absolute Percentage Error

ML Machine Learning

MLF Micro Learning Factory

MR Multivariate regression

MSE Mean Squared Error

NOCT Nominal Operating Cell Temperature

NWP Numerical Weather Prediction

OPDS Open Data Soft

PV Photovoltaic

PSE Prevention, health, and environment

RF Random Forest

ROI Return on Investment

RTE Réseau de Transport d'Electricité

ROS Robot Operating System

SNBC National Low-Carbon Strategy

STC Standard Test Conditions

SVM Support Vector Machine

SVR Support Vector Regression

SVMR Support Vector Machine Regression

TWh Terawatt-hours

VR Virtual Reality

XGBR Extreme Gradient Boosting Regressor

Table of Contents

Introduction

The increasing global demand for electricity, coupled with the urgent need to mitigate climate change, has spurred a rapid transition towards renewable energy sources. Solar photovoltaic (PV) systems, particularly building-integrated photovoltaics (BIPV), have emerged as a promising solution to decarbonize the building sector and contribute to a more sustainable energy future. Accurate prediction of energy production from PV systems is crucial for optimizing system design, maximizing energy yield, and assessing the economic viability of solar energy projects. This report details the development of an AI-powered tool designed to accurately predict the energy production of BIPV systems, contributing to the decarbonization of real estate parcs and supporting informed decision-making for stakeholders.

This report is structured as follows: Chapter 1 outlines the partnership and project overview, providing context and motivation for the development of the AI tool. Chapter 2 presents a comprehensive literature review, exploring existing research on renewable energy, solar energy, and the application of AI in energy production predictions. Chapter 3 focuses on reproducing and analyzing the existing approach developed by Youssef & Ilyas, highlighting its limitations and areas for improvement. Chapter 4 details our novel approach, encompassing the generation of a synthetic dataset using BimSolar, the selection of machine learning models, the integration of carbon emission tracking, and the evaluation of model performance. Chapter 5 showcases the development of a user-friendly web application that integrates the trained models, providing an intuitive interface for data generation, energy prediction, and decarbonization assessment. Finally, Chapter 6 summarizes the findings, discusses conclusions from the current work, and outlines future prospects for research and development.

Chapter 1

Partnership and project overview

This chapter is divided into two main parts. The first part presents the context of the project, the motivation behind it, and the objectives aimed to be achieved. The second part introduces the CESI LINEACT laboratory as well as the CCCA-BTP company.

1.1 Project overview

1.1.1 Context and motivation

Electricity is central to the functioning of modern societies and economies, and its importance is continuously amplified as electrically-powered technologies gain widespread adoption. Although power generation stands as the primary contributor to global Carbon dioxide (CO2) emissions, this sector is leading the transition towards net-zero emissions through the rapid growth of renewable energy sources like solar and wind power. Simultaneously ensuring reliable and affordable electricity access for consumers while mitigating global CO2 emissions presents a central challenge in the ongoing energy transition.

The International Energy Agency (IEA) forecasts a more rapid rise in global electricity demand over the next three years, projecting an average annual growth rate of 3.4% through 2026. This increase in energy consumption is fuelled by the remarkable expansion of the data center sector and the increasing electrification of residential and transportation sectors. According to the IEA report [1], electricity consumption from AI, data centers, and cryptocurrency operations is going to double by 2026.

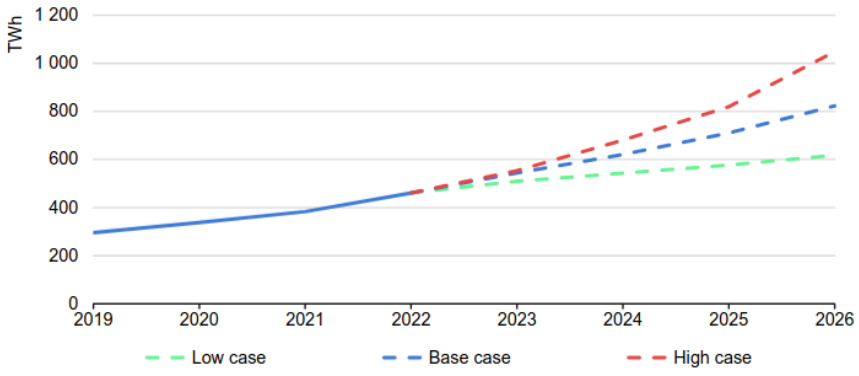

Figure 1: Global electricity demand from data centres, AI, and cryptocurrencies, 2019-2026.

Data centers alone consumed a staggering 460 Terawatt-hours (TWh) in 2022, and by 2026, their consumption is estimated to surpass 1000 TWh, equivalent to the entire

electricity consumption of Japan. Figure 1 illustrates that the electricity demand kept increasing from 2019 to 2026.

The Fit for 55 package [2], a European Union initiative aimed at reducing greenhouse gas emissions by at least 55% by 2030, highlights the significant impact of buildings on energy consumption and emissions. Within the European Union (EU), buildings account for a staggering 40% of final energy consumption and contribute 36% of energy-related greenhouse gas emissions. To address this challenge, the package requires the installation of solar technologies on all new buildings by 2028, where technically feasible and economically viable. Furthermore, residential buildings undergoing major renovations will be required to comply with this directive by 2032. This measure holds profound significance, as buildings capable of generating their own energy to meet consumption demands will substantially reduce reliance on grid-supplied energy, consequently decreasing CO2 emissions.

Figure 2: The sectors covered by the Fit for 55 package [2].

Figure 2 illustrates the sectors covered by the Fit for 55 package. The package encompasses reforms in the emissions trading system, transportation, buildings, agriculture, waste management, land use, forestry, and energy taxation. Specific measures include toughening CO2 emission standards for vehicles, increasing the uptake of greener fuels in aviation and maritime sectors, regulating methane emissions, boosting renewable energy sources, and making buildings more energy-efficient. The package also aims to address emissions outside the EU and support citizens and businesses affected by the transition through a dedicated fund. Additionally, it focuses on promoting sustainable transport, achieving climate goals in land use and forestry sectors, and shifting from fossil gas to renewable and low-carbon gases. Overall, the Fit for 55 plan is a multifaceted approach to tackle climate change and pave the way for a greener, more sustainable future for the EU. and energy.

According to the Réseau de Transport d'Electricité (RTE) [3], the French strategy for the future to reach net-zero by 2050, known as the National Low-Carbon Strategy (SNBC), involves reducing energy consumption by 40% in the upcoming 30 years, from 1600 TWh to 930 TWh in 2050. The strategy, which is re-evaluated every five years with the latest version published in 2020, outlines the reference framework for RTE's 2050 Energy Futures, examining numerous scenarios that all comply with the carbon neutrality target by 2050. The strategy mentions that the energy will be produced primarily from renewable

energy sources and biomass, while reducing the use of non-renewable energies.

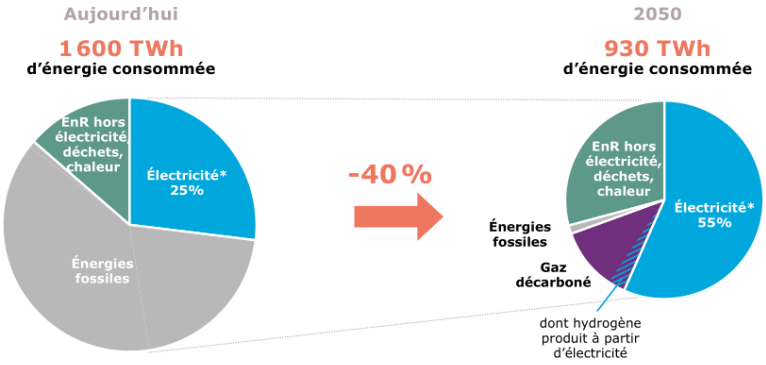

Figure 3: The French strategy for the future to reach net-zero by 2050 [3].

The growing demand for electricity, driven by the expansion of data centers, AI, and the electrification of various sectors, emphasizes the urgent need for sustainable energy solutions. Simultaneously, initiatives such as the EU's Fit for 55 package and France's National Low-Carbon Strategy (SNBC) emphasize the imperative to reduce greenhouse gas emissions and foster renewable energy sources. In this context, the development of an Artificial Intelligence tool capable of accurately predicting and forecasting the energy generated by PV panels emerges as an important project. By providing reliable estimates of the energy production potential from solar panels, this tool can play an important role in optimizing the integration of renewable energy sources into buildings, thereby mitigating the environmental impact of electricity generation and accelerating the transition towards a sustainable future.

1.1.2 Objectives

This project aims to develop a user-friendly tool that helps building owners and stakeholders understand and optimize the potential of solar energy for their buildings. The ultimate goal is to contribute to the decarbonization of CCCA-BTP's real estate parc by supporting the integration of PV systems.

To achieve this objective, the first step is to reimplement the work done by Youssef & Ilyas in their research [5]. This is an important step because it helps to better understand the problem, the solution they proposed, and to develop our own solution.

After reimplementing their work and pinpointing the areas that need improvement, the focus shifts to developing our own solution. Various ways to generate synthetic data to enrich our dataset will be explored, and the impact of training the models on the environment will be assessed. Additionally, efforts will be made to automate most of the steps in the pipeline.

After developing the model, its performance will be evaluated on some study cases. Three buildings will be used in the study case: the CESI building in Strasbourg, the Campus BTP CFA OCCITANIE in Toulouse-Muret, and the Association ouvrière des Compagnons du Devoir et du Tour de France (AOCDTF) building in Strasbourg.

Finally, the last objective is the development of a web application that will allow users to generate predictions for their buildings. Evaluate the energy performance of the building,

the emission performance of the building, as well as the decarbonization rate.

This internship was generously financed by the Comité de concertation et de coordination de l'apprentissage du bâtiment et des travaux publics (CCCA-BTP), reflecting their strong commitment to promoting innovation and sustainable practices within the construction and public works sectors. Their support for this project underscores their dedication to exploring the potential of AI and renewable energy in shaping a more sustainable future for the industry.

1.2 CCCA-BTP

CCCA-BTP is a national association based in France, dedicated to promoting and coordinating vocational training in the construction and public works sectors. Established as a collaborative effort between employer federations and employee unions, the organization aims to enhance the quality of professional training, particularly through apprenticeships, for young people in the building and public works industries.

Figure 4: The CCCA-BTP logo.

1.2.1 Missions and vision

The CCCA-BTP is dedicated to advancing vocational training, particularly through apprenticeships, in the building and public works sectors (BTP). The organization embodies the commitment of professionals within these sectors to the education and professional development of young people. By establishing various agreements, social partners have expressed their intent to ensure that every young individual involved in BTP apprenticeships becomes a qualified professional and secures sustainable employment within the industry.

Key objectives and actions of the CCCA-BTP include conducting thorough surveillance across all BTP and professional training domains, producing barometers and territorial studies in collaboration with the OPMQ (Observatoires prospectifs des métiers et des qualifications), designing educational directions and resources, promoting the construction sector and its careers, raising awareness about the implementation of quality management and certification systems, organizing expert-led thematic meetings, leading and coordinating training organization consortia to respond to national and European project calls, and innovating within BTP professions and training methodologies, supporting training organizations with funding for experimental and innovative projects.

The CCCA-BTP boasts expertise in training engineering, with nearly 80 years of experience as a pedagogical expert in alternance training. It defines educational orientations, provides educational resources, creates digital solutions for pedagogical practices, and contributes to the development of competency frameworks for BTP diplomas.

In terms of European and international projects, the CCCA-BTP engages in European mobility and strategic partnerships, benefiting 2,000 individuals with European mobility actions since 2016 and developing 300 professional patents with European components. It forms strategic European partnerships to foster innovative pedagogy, including membership in the transnational European network REFORME. These initiatives highlight the organization's commitment to enhancing vocational training on a broader scale, ensuring its alignment with evolving European standards and methodologies.

Innovation is a driving force for the CCCA-BTP, with its Innovation Laboratory (Winlab') acting as a project accelerator. Winlab' collaborates with major training actors on Building Information Modeling (BIM), virtual reality, distance learning, and more, promoting innovative solutions deployed by training organizations and providing funding to support experimentation and innovation through calls for projects.

1.2.2 Digital learning

The CCCA-BTP offers training organizations for construction professions solutions it has designed to develop the integration of digital technology into educational practices, the promotion of work-study pedagogy, as well as socio-educational support for apprentices. Let's talk about two solutions, C'KI L'BOSS and APTYCE.

C KI L'BOSS is a serious game designed to introduce learners to the world of entrepreneurship. C KI L'BOSS is a fun and interactive serious game application that combines both management and strategy to raise awareness among construction training learners about the entrepreneurial spirit, business management, and creation. C KI L'BOSS allows construction training learners to immerse themselves in managing a company and to become familiar with key entrepreneurial principles: sustainability, cash flow, management, team leadership, and reputation.

Figure 5: An overview of the C KI L'BOSS game.

C KI L'BOSS is a game designed to :

- Meet the standards of construction trades training diplomas and the expectations of industry professionals in terms of business creation or takeover.

- Incorporate Prevention, health, and environment (PSE) programs.
- Contribute to the development of learners' autonomy and collaborative spirit.
- Be accessible on all types of devices (computers, tablets, and smartphones).

Aptyce is an open and distance learning platform provided to training organizations in the construction trades. It allows learners to access digital learning programs created and led by their Centre de formation d'apprentis (CFA) instructors. CFAs create their own training programs in Aptyce based on their own modules or the modules made available to them. Aptyce includes a module design tool that allows for the creation of training programs for groups of learners, and even for each individual learner, enabling collaborative and interactive work, as well as tracking learners' progress through the programs.

Figure 6: The Aptyce logo.

Apstyce has the following features :

- Ease of use thanks to its ergonomic design and features that require no technical computing skills.
- Enables collaborative work. Trainers, as authors on Aptyce, create their content and can share it.
- Accessible to users via a computer or a mobile application on a smartphone, available 24/7.
- Offers activities that allow for the creation of multimodal programs for learners.

1.2.3 Innovation inside CCCA-BTP

CCCA-BTP created the WinLab' in 2017, signaling a significant move to promote innovation within the BTP sector. This initiative not only demonstrates a commitment to adapting to the evolving needs of the construction industry but also highlights the recognition of the crucial role startups and emerging technologies play in this landscape. By embracing innovation and providing a platform for startups to thrive, organizations like CCCA-BTP contribute to the ecosystem's vitality and resilience. Such initiatives not only stimulate creativity and experimentation but also facilitate the integration of cutting-edge technologies and methodologies into training programs, ensuring that future professionals are equipped with the skills necessary to tackle the challenges of tomorrow's construction industry.

Figure 7: The Winlab' logo.

To fulfil this mission on a daily basis, the WinLab' is characterized by its central role as a trend scout: a true "off-site innovation laboratory" that supports training organizations and construction sector companies in implementing innovative solutions, offering them resources and advice to develop their innovative projects. It provides startups and innovative companies with opportunities to present their solutions and benefit from collaborative workspaces with key industry players.

Here are three examples of startups currently being supported by WinLab':

Figure 8: The logos of the three startups supported by WinLab'.

1. **Ubeton:** Ubeton is revolutionizing concrete usage in construction projects through its innovative platform. By enabling artisans to efficiently create concrete pouring projects and accessing surplus concrete volumes from nearby construction sites, Ubeton optimizes resource utilization and reduces costs for both artisans and concrete plants. Ubeton gains access to resources and collaborative opportunities to further enhance its platform and expand its impact in the construction industry.

2. **Carbon Saver:** Carbon Saver offers a SaaS (Software as a Service) software dedicated to eco-design, aiding in the selection of environmentally friendly materials and practices during the design phase of construction projects. Their flagship product, Bat'impact, provides users with an intuitive environmental score system and actionable improvement recommendations.

3. **Renalto:** Renalto provides a comprehensive solution for building renovation professionals, streamlining business management and integrating energy renovation seamlessly into their workflow. Their user-friendly application incorporates the best available technologies to accelerate energy renovation efforts.

These companies exemplify the diverse range of innovative solutions supported by WinLab', contributing to the advancement and sustainability of the construction industry.

1.3 CESI LINEACT laboratory

CESI LINEACT was established in 2015 as part of a strategy to structure CESI's research activities in service of its educational programs, and was designated as a Host Team (now Research Unit) in 2018. CESI LINEACT has built its positioning around CESI's training fields, conducting research activities that serve the domains of industry, sustainable cities, digital services, and future training. The unit has positioned itself at the interface of physical and digital worlds with a transdisciplinary SHS-STIC approach and offers a vision of applied research that places the uses and users of technological systems at the center of the technology's purpose, whether for economic, cultural, or societal aims, such as the training of professionals or future professionals.

Figure 9: The CESI LINEACT logo.

CESI LINEACT supports the needs of CESI's training programs through new training methodologies, the production of new knowledge, and their integration into educational content, as well as through the development and use of dual-purpose platforms for training and research. The unit also supports the transformation of CESI's training programs.

1.3.1 Organization

CESI LINEACT is the research laboratory of the CESI group as we said before, with activities implemented across 23 sites in France, grouped into 6 regions or campuses. As of March, 2024, it comprises 98 associate professors and 57 PhD students, as well as 12 research and innovation engineers, 2 administrative and financial executives and 2 assistants.

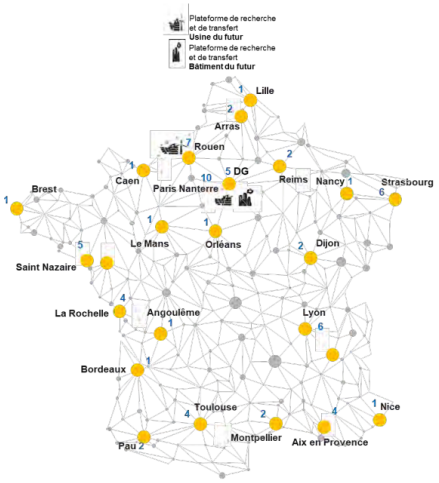

Figure 10: Distribution of CESI teacher-researchers by campus.

Now, let's describe the organizational structure at CESI LINEACT.

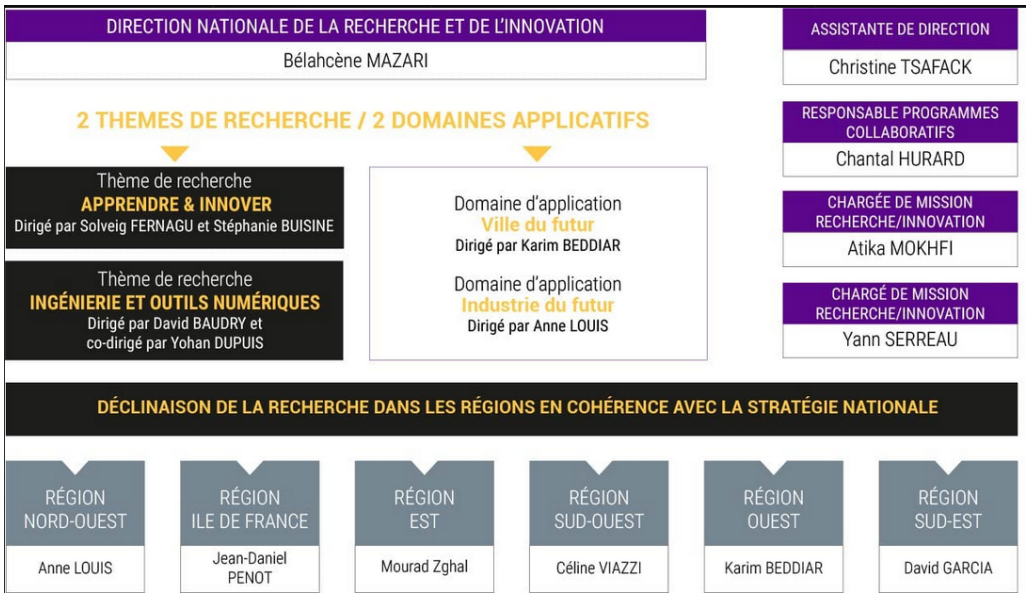

Figure 11: CESI LINEACT organizational chart.

CESI LINEACT is organized with a National Research and Innovation Directorate in charge. This department is led by Bélahcène MAZARI and supported by Christine TSAFACK. Atika MOKHFI and Yann SERREAU act as Research/Innovation Mission Officers, while Chantal HURARD manages Collaborative Programs.

The core research activities are divided into two main teams. The first, focusing on the "City of the Future" application domain, is the Learning & Innovation Research Team led by Solveig FERNAGU and Stéphanie BUISINE. The other, concentrating on the "Industry of the Future" domain, is the Engineering & Digital Tools Research Team co-directed by David BAUDRY and Yohan DUPUIS.

CESI LINEACT operates with a decentralized structure, meaning regional teams work alongside the national strategy. These regional teams are led by: Anne LOUIS (Nord-Ouest), Jean-Daniel RENOT (Ile de France), Mourad ZGHAL (Est), Céline VIAZZI (Sud-Ouest), Karim BEDDIAR (Ouest), and Dr. Saeed MIAN QAISAR (Sud-Est, as of January 4, 2023).

1.3.2 Research teams

The specificity of CESI LINEACT lies in organizing its research according to two interdisciplinary scientific teams (learning and innovation, engineering and digital tools) and two application domains. These focus on the profound and increasingly rapid transformations of industry and cities.

- **Learning and innovation:** This team primarily involves Cognitive Sciences, Social Sciences, and Educational Sciences with the goal of studying, designing, and evaluating learning and innovation ecosystems in education and work, considering their socio-technical, economic, and human dimensions.

- **Engineering and digital tools:** This team focuses on scientific domains that include Digital Sciences, Industrial Systems Engineering, Operations Research, and Engineering Sciences.

Both teams develop and cross their research in the three application domains of Industry 5.0, Construction 4.0, and Sustainable City, as well as Digital Services. Three research and transfer platforms, whose size, activity volume, and digital duplication allow them to shine across all CESI campuses, support all of LINEACT's work: two are dedicated to the factory of the future at the Rouen and Nanterre centers, and one to the building of the future at the Nanterre center.

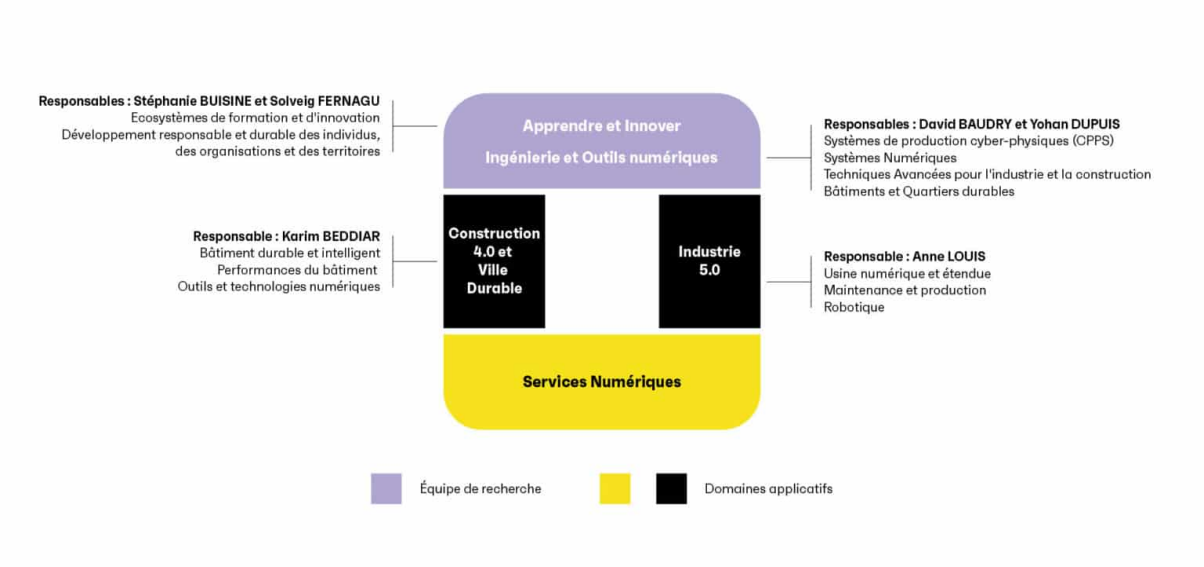

Figure 12: Organisation of CESI LINEACT research.

The research conducted at CESI LINEACT is structured around the following axes:

- High-quality research.
- Research with and for society.
- Applied and integrative research at the intersection of digital and physical worlds, serving territorial specificities.

1.3.3 Technological platforms

CESI LINEACT has state-of-the-art platforms and equipment, as well as a technical team composed of research engineers and a competent administrative unit to manage them. CESI LINEACT's equipment enables it to offer innovative and high-quality solutions and to undertake ambitious projects. Let's list some of the platforms and equipment available at CESI LINEACT.

1.3.3.1 Platform dedicated to the industry of the future

The "Industry of the Future" platform, located on the CESI campus in Rouen, replicates a modular production workshop featuring an automated production line, industrial manual workstations for assembly and quality control operations, AGVs (Automatic Guided

Vehicles) for transporting parts and products, manipulators, cobots for collaborative assembly tasks, Industrial Internet of Things (IIoT) for data exchange, and Human-Machine Interfaces based on augmented reality or virtual reality to assist or train in assembly or maintenance operations. This platform is complemented by a digital twin developed within the laboratory's research activities and is integrated into various educational projects.

1.3.3.2 AR/VR equipment integrated with digital twins

Augmented Reality (AR) and Virtual Reality (VR) are immersive technologies, offering numerous opportunities for research and industry. In research, AR and VR, often coupled with digital twins, simulate complex environments to reduce development costs and time, minimize design errors and defects, and visualize complex data in an interactive and intuitive environment. These technologies enable the exploration and testing of research hypotheses or the more effective and interactive training of professionals, especially with the advent of Industry 5.0 and the reintegration of humans into industrial processes. AR and VR are also used for training and simulation in complex industrial environments and dangerous or difficult situations, such as interventions in ATEX areas (explosive atmospheres in the presence of an ignition source).

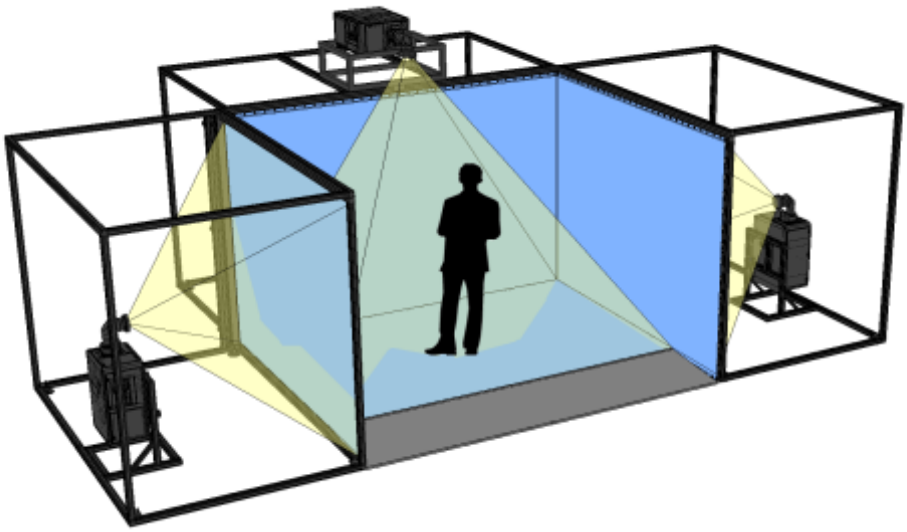

Figure 13: An illustration of the CAVE system. Image credit: Visbox, Inc. [4].

Most CESI campuses have AR/VR headset equipment. One campus features a dedicated technical platform allowing collaborative situations with multiple users in VR, a 4-sided Cave Automatic Virtual Environment (CAVE), data acquisition and processing systems to study human behavior (motion capture, physiological data, etc.).

1.3.3.3 Robotic platforms

Initially launched by a few CESI centers, robotics is rapidly expanding at CESI with the arrival of 18 robots distributed across various campuses. In research, robotics enables the performance of complex experiments, exploration of dangerous or inaccessible

environments, and development of new technologies and applications. Robotics can improve the efficiency and productivity of manufacturing processes, reduce production costs and time, and enhance product quality. Robots are used in numerous industrial sectors such as automotive, aerospace, electronics, logistics, and healthcare to perform repetitive, dangerous, or highly precise tasks.

Additionally, robotics offers great potential for innovation and the development of new applications. Collaborative robots, for example, are designed to work alongside humans, opening new possibilities for task automation in many industrial fields. Robotic platforms can also address common themes among the research teams "Learning and Innovating" and "Engineering and Digital Tools," particularly regarding Human-robot interactions.

Figure 14: The MLF in Strasbourg.

The figure above shows the MLF, a robotic platform that allows students to learn about the principles of Industry 4.0 and the operation of a production line. The MLF is a modular and scalable platform that can be adapted to various training needs and is equipped with a conveyor belt, a robot arm, and a vision system.

1.3.4 Research projects

CESI LINEACT is involved in numerous research projects that contribute to the advancement of knowledge and the development of innovative solutions in various fields. The laboratory's research activities are structured around three main axes: Industry 5.0, Construction 4.0, and Sustainable City. Let's discuss some of the research projects conducted by CESI LINEACT.

1.3.4.1 JENII

The Jumeaux d'Enseignement Numériques, Immersifs et Interactifs (JENII) project, supported by the "DemoES" 2021 investment program, focuses on the development and optimization of Digital Twins (JNs) in higher education. This project aims to identify and

support educational institutions ready to demonstrate all dimensions of digital transformation, including pedagogy, equipment, and usage. With a significant budget of €14 million, including €1.75 million allocated to CESI, JENII involves a consortium comprising ENSAM, CNAM, CEA Tech, and CESI. The project emphasizes the creation of hybrid, multimodal, and individualized learning paths and the R&D of Digital Twins in advanced technological, pedagogical, and usage dimensions.

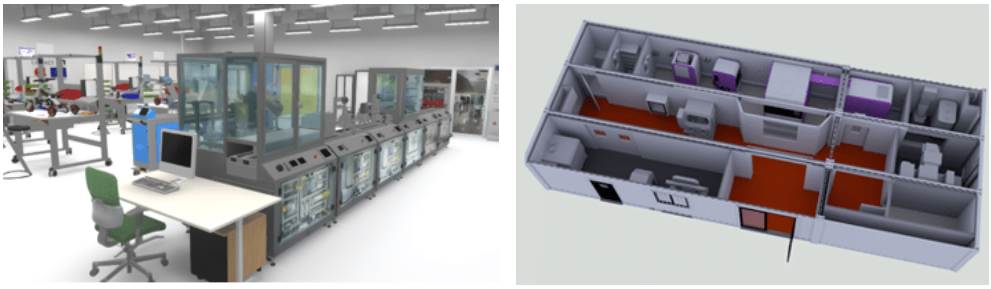

Figure 15: JENII, The digital twins of CESI platforms.

1.3.4.2 Tiago++ robots

CESI has acquired six Tiago++ robots to support its pedagogical and research activities. These humanoid robots are equipped with various sensors and actuators, including a lidar, RGB-D camera, stereo microphones, and robotic arms, allowing them to interact naturally with students and researchers. Tiago++ serves as both an educational tool, guiding students through practical demonstrations, and a research instrument, enabling the exploration of advanced concepts in robotics and human-robot interaction. The robots' integration with the Robot Operating System (ROS) facilitates the development of standardized robotic modules that can be shared across the scientific community, fostering collaborative innovation.

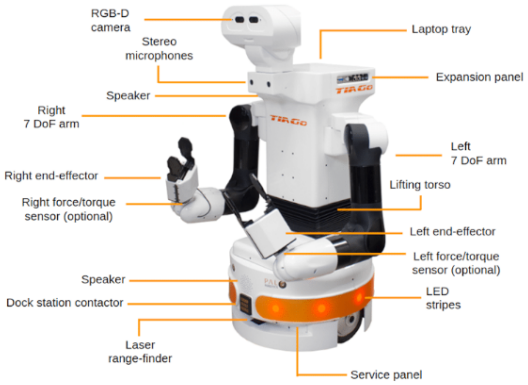

Figure 16: The Tiago++ robot.

1.4 Conclusion

In this chapter, we have presented the context of our project, focusing on the importance of renewable energy sources, particularly solar energy, in the ongoing energy transition. We have highlighted the growing demand for electricity, driven by the expansion of data centers, AI, and the electrification of various sectors, and the urgent need for sustainable energy solutions. Initiatives such as the EU's Fit for 55 package and France's National Low-Carbon Strategy (SNBC) underscore the imperative to reduce greenhouse gas emissions and foster renewable energy sources. We have also introduced the CCCA-BTP, a national association dedicated to promoting vocational training in the construction and public works sectors, and the CESI LINEACT laboratory, which conducts research activities in the fields of industry, sustainable cities, digital services, and future training. We have discussed the organizational structure of CESI LINEACT, its research teams, technological platforms, and ongoing research projects. The following chapters explore the development of a machine learning or deep learning model for predicting energy production in building-integrated PV systems, aiming to support the transition towards sustainable energy solutions.

Chapter 2

Literature review

In this chapter, a literature review of the research conducted in the field of renewable energy, particularly solar energy, and the use of machine learning and deep learning techniques to predict the energy production of PV systems will be presented. The importance of accurate energy production predictions for PV systems, the challenges associated with these predictions, and the various approaches proposed in the literature to address these challenges will be discussed. Additionally, the potential of synthetic data generation to enhance the performance of machine learning models and the impact of training these models on the environment will be explored.

2.1 Renewable energy and solar energy

Renewable energy sources, such as solar, wind, hydro, and geothermal energy, play a crucial role in the ongoing energy transition towards a sustainable and low-carbon future. Solar energy, in particular, has gained significant attention due to its abundance, accessibility, and environmental benefits. Solar PV systems, which convert sunlight into electricity, have become increasingly popular for residential, commercial, and industrial applications. The integration of PV systems into buildings, known as Building-Integrated Photovoltaics (BIPV), offers a promising solution to reduce energy consumption, lower greenhouse gas emissions (GHG), and enhance the sustainability of the built environment.

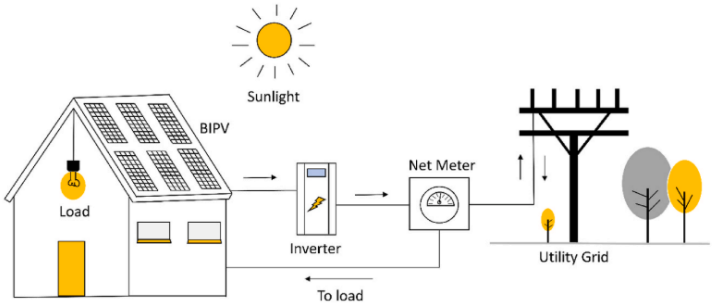

Figure 17: Building-Integrated Photovoltaics (BIPV) system on a house [16].

Figure 17 illustrates a BIPV system installed on a house, showcasing the integration of solar panels into the building's facade and roof. BIPV systems not only generate clean and renewable electricity but also contribute to the architectural design, energy efficiency, and sustainability of the building. The panels are connected to an inverter that converts the Direct current (DC) electricity generated by the solar cells into Alternating current (AC) electricity, which can be used to power appliances, lighting, and other electrical devices

in the building. In the case of excess electricity production, the surplus energy can be stored in batteries or fed back into the grid, depending on the system configuration and local regulations.

2.1.1 Importance of energy production predictions

Accurate predictions of energy production from PV systems are essential for optimizing their performance, maximizing energy yield, and ensuring the economic viability of solar energy projects. Energy production predictions help stakeholders, such as building owners, energy managers, and grid operators, to estimate the amount of electricity generated by the PV system over a specific period, typically daily, monthly, or yearly. These predictions are crucial for planning and decision-making, including system design, sizing, installation, maintenance, and financing.

Moreover, energy production predictions enable building owners to assess the Return on Investment (ROI) of the PV system, evaluate its environmental impact, and compare different system configurations or technologies.

2.1.2 Challenges in energy production predictions

Predicting the energy production of PV systems is a complex task due to various factors that influence solar energy generation, such as weather conditions, shading effects, system orientation, tilt angle, module efficiency, and the different types of losses in the system. The variability and uncertainty of these factors pose significant challenges for accurate energy production predictions, requiring sophisticated modeling techniques and data-driven approaches to address them effectively.

Figure 18: Key losses in a PV system [17].

Some of the key challenges in energy production predictions for PV systems include:

- **Shading effects**, which can reduce the efficiency of the PV system due to interruption of direct sunlight hitting the panels.
- **Dust effects**, which can decrease energy output and efficiency, especially in dusty environments.
- **Solar cell array optical losses**, where light is reflected off the panel surface instead of being absorbed and interacting with the electrons.
- **Solar cells spectral responses**, as they do not utilize all wavelengths of sunlight equally.
- **Solar energy system loss due to irradiance level**, as energy production is not linear with decreasing irradiance levels.
- **Solar panel cells thermal loss**, where the solar cell output decreases with increasing temperature above the standard test conditions.
- **Solar PV modules mismatch loss**, arising from variations in electrical properties among solar modules in an array.
- **Solar panels DC wiring losses**, due to the resistance of cables causing voltage drops and power losses as heat.
- **Losses in solar power inverters**, as the efficiency of inverters is typically around 97%, leading to some power loss during conversion.
- **Energy system voltage drop**, caused by resistance in the wiring run, affecting the voltage supplied to the inverter.

To overcome these challenges and improve the accuracy of energy production predictions, researchers have explored various modeling techniques, including physical models, statistical models, Machine Learning (ML), and Deep Learning (DL) algorithms. Let's delve into the literature to review the different approaches proposed to predict the energy production of PV systems.

2.2 Applications of AI in energy production predictions

ML and DL techniques have gained popularity in recent years for predicting the energy production of PV systems. ML and DL algorithms offer powerful tools to analyze complex data, identify patterns, and make accurate predictions based on historical and real-time information. These techniques leverage the computational power of computers to process large datasets, extract meaningful features, and train predictive models that can estimate the energy output of PV systems under different conditions.

Alomari et al. [18] proposes an efficient forecasting model to predict the next-day PV power using the Levenberg-Marquardt (LM) and Bayesian Regularization (BR) algorithms along with real-time weather data. The aim is to improve the stability of power production in grid-connected PV power plants, which is crucial for controlling and operating the electrical grid. The study investigates the correlations between global solar

irradiance, temperature, solar PV power, and time of the year to extract knowledge from historical data for developing a real-time prediction system.

The researchers collected a two-year dataset (May 2015 to May 2017) containing 17,544 weather records and 17,544 PV power values from the Applied Science Private University (ASU) in Amman, Jordan. The dataset was filtered to ensure consistency by removing any missing records or mismatched weather and PV power data. The data was normalized between 0 and 1 for homogeneity and reliable machine learning experiments.

The study employed Artificial Neural Networks (ANNs) as a powerful machine learning technique for mapping non-linear inputs through adjustable weights into desired targets.

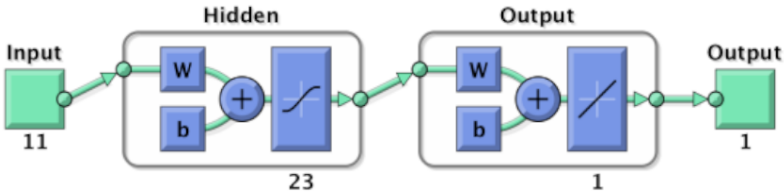

Figure 19: The proposed forecasting model by Alomari et al. [18].

The authors conclude that the proposed predictive forecasting model, by applying the LM and BR algorithms to neural networks and correlating historical weather data with photovoltaic outputs, can provide excellent PV power forecasts for the next 24 hours. These predictions can be valuable for energy management systems and power control systems of grid-tied PV plants. The authors plan to develop the proposed model into a real-time online application in their future work.

Bowoo et al. [19] introduces a lightweight hybrid model for predicting solar photovoltaic (PV) generation on an hourly basis, utilizing both remote sensing data from satellite images and Numerical Weather Prediction (NWP) data. The model combines spatial features extracted from infrared satellite imagery with temporal data from hourly weather datasets to capture the spatio-temporal characteristics impacting solar PV generation. It defines regions of interest (ROIs) within the satellite images corresponding to the solar plant location and surrounding areas to reduce computational load while focusing on relevant spatial information.

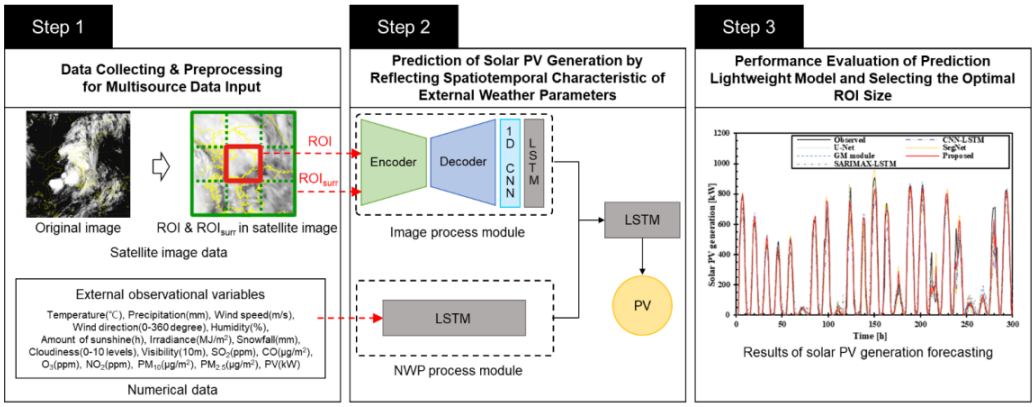

Figure 20: The research framework proposed by Bowoo et al. [19].

The research framework proposed by Bowoo et al. [19] is illustrated in Figure 20. The model has two key modules - an image processing module that uses a CNN-LSTM ensemble to predict cloud and particulate matter movement from satellite images, and an NWP module that processes numerical weather data. The output from both modules is combined to make the solar prediction, accounting for both spatial cloud patterns and temporal weather variations.

The paper demonstrates the effectiveness of integrating multi-source remote sensing and NWP data and using focused ROIs within satellite imagery for accurate yet lightweight solar forecasting useful for smart grid operations.

Luo et al. [20] proposes a machine learning-based multi-objective prediction framework for simultaneously predicting multiple building energy loads and electrical power production from building-integrated photovoltaics (BIPV). The motivation is that accurate day-ahead prediction of heating, cooling, lighting loads, and BIPV power production is essential for effective building energy management, given the complex and variable nature of these loads due to changing weather conditions and building characteristics.

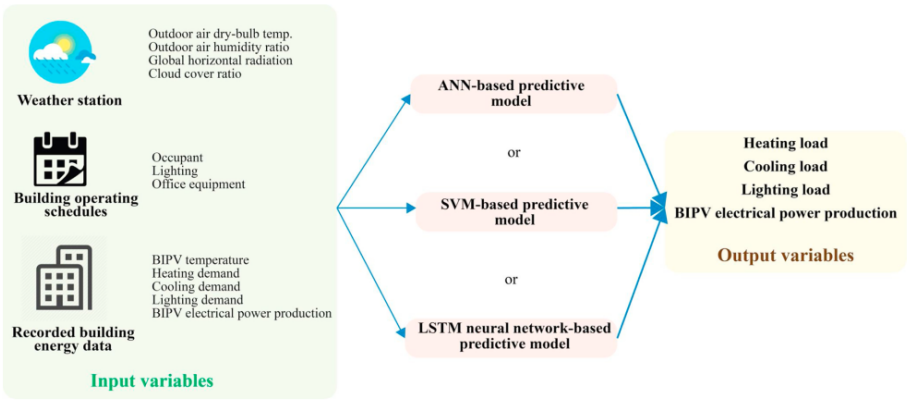

Figure 21: The schematic diagram of the proposed multi-objective prediction framework. [20].

The paper proposes three machine learning techniques - Artificial Neural Networks (ANN), Support Vector Regression (SVR), and Long Short-Term Memory (LSTM) - for this multi-objective prediction task. Since heating, cooling, lighting loads and BIPV power production share common affecting factors like weather data and building operations, using a multi-objective framework to predict them simultaneously can save computational time compared to separate models.

The study evaluates the proposed predictive models on a reference office building with BIPV and lighting control systems. Different building material properties (wall thermal conductivity, window thermal properties, window-to-wall ratio) are investigated to test the robustness of the framework. A validated TRNSYS building simulation model is used to generate the training data based on weather data from Heathrow Airport and pre-set operating schedules.

The results show that the ANN-based model achieved the smallest mean absolute percentage error, while the SVM-based model had the shortest computation time. The proposed multi-objective prediction framework is deemed valuable for various building

energy management tasks like demand-side management, control strategies, and fault detection by providing day-ahead forecasts of loads and generation.

The authors suggest extending the framework to other building types like hotels, residential, and hospitals by modifying the input variables in the database according to the characteristics and usage patterns of those buildings. The accurate multi-objective prediction of energy loads and BIPV power aids in effective utilization of solar energy and daylighting for reducing overall energy consumption in buildings.

Pamela et al. [21] present a new technique for forecasting the 24-hour ahead stochastic energy output of photovoltaic (PV) systems based on daily weather forecasts. The motivation is that accurately predicting PV output is essential for managing the operation and economic performance of power systems that utilize intermittent solar energy. The researchers compared the performance of their hybrid technique against conventional linear regression and artificial neural network models.

The study used data from an operational PV system in Newquay, Cornwall consisting of 16 solar panels. The dependent variable was the energy output of the PV system, while the independent variables were average daily meteorological factors like temperature, pressure, humidity, rainfall, solar irradiance, wind direction and speed. Data from 2012 to 2013 was collected, with the first 500 samples used for training and the remaining 193 for testing the models.

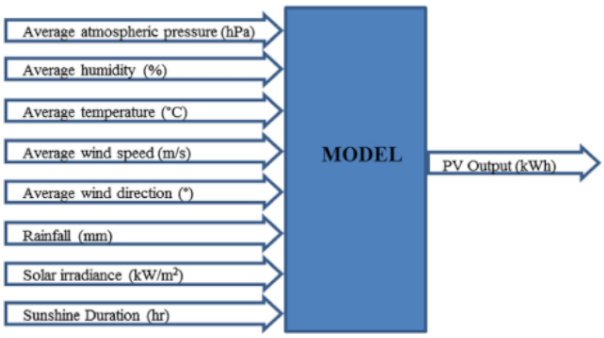

Figure 22: The independent and dependent variables. [21].

Three single-stage models were initially designed - Generalized Regression Neural Network (GRNN), Feedforward neural network (FFNN) and multiple linear regression (MLR). The models were evaluated using metrics like root mean squared error (RMSE), Mean Absolute Error (MAE), mean bias error (MBE), and correlation coefficient (R). The hybrid models performed better than their single-stage counterparts.

The researchers concluded that accurate forecasting of PV system power output is critical for proper planning and operation of power systems. Their proposed hybrid approach using stepwise regression to select the most relevant meteorological parameters as inputs to neural network models showed improved prediction accuracy compared to single-stage models. The hybrid models, especially the SR-FFNN, were able to forecast the 24-hour ahead daily energy production very close to the measured values from the operational PV

system. Given their simpler architectures, the hybrid models are capable tools for PV energy prediction, particularly in locations where limited meteorological data is available.

Abhishek et al. [22] compared the performance of three machine learning techniques - Multivariate regression (MR), Support Vector Machine Regression (SVMR), and Gaussian regression (GR) - for predicting the power output of solar PV panels. The researchers investigated how environmental factors like solar radiation, ambient temperature, and relative humidity impact PV panel output. The results showed that SVMR models had superior predictive capabilities compared to GR and MR models.

The researchers collected data over 60 days (January-March 2022) from a solar installation with two 3.2 kW PV systems at Aditya Engineering College in India. Environmental parameters like temperature, humidity, solar radiation, and output power were recorded daily at noon. The study period was chosen to minimize rainfall's influence. The data captured atmospheric variations' effects on daily power output, aiming to enhance the prediction model's accuracy.

The selection of SVMR, GR, and MR techniques was based on their ability to handle complex datasets, model nonlinear relationships, quantify prediction uncertainty, and incorporate multiple predictors simultaneously. These techniques are well-established, computationally efficient, and have been validated in various domains, including solar energy prediction.

The study's findings reveal that the SVMR algorithm emerges as the top-performing model for predicting solar PV panel output power, with an exceptional R^2 value of 0.99 and minimal errors. The research provides practical guidance for optimizing solar power plant performance and highlights the efficacy of machine learning techniques in solar energy generation, potentially facilitating wider adoption of sustainable energy practices.

2.3 Limitations and improvements

While the research presented by Alomari et al. [18], Bowoo et al. [19], Luo et al. [20], Pamela et al. [21], and Abhishek et al. [22] has made significant contributions to the field of solar PV energy prediction, our work aims to address some of their limitations.

Firstly, most of the studies relied on limited datasets, often specific to a single location or a small number of PV systems. This limits the generalization of their findings and the applicability of their models to different contexts. Our work addresses this limitation by generating a much larger and more diverse dataset using BimSolar, incorporating data from various locations across France, a wider range of solar panels, and diverse building types. This approach significantly expands the model's ability to generalize to unseen data and improves its robustness.

Secondly, the existing research often focused solely on meteorological factors and PV system characteristics. They often neglected incorporating building information, which can significantly influence energy production. We integrate BIM data into our dataset, providing crucial information about the building, such as the surface area and location for PV panel installation, as well as the exploitation rate. This allows the model to better understand the building's energy production potential.

Thirdly, most of the research focused on forecasting the energy production for a specific

period in the future. This approach relies on historical data and may not be applicable to new PV systems or scenarios with limited historical data. We shifted our focus from forecasting to prediction, enabling us to directly estimate the energy production for a given day based on current conditions without relying on historical data. This approach is more flexible and adaptable to new scenarios.

Fourthly, the existing research often lacked an in-depth analysis of the energy consumption and carbon emissions associated with model training. This is crucial for developing sustainable AI solutions. We integrated the CodeCarbon library to measure the energy consumption and carbon emissions during model training. This allows us to select energy-efficient models, minimizing the environmental impact of our work.

Finally, most of the research focused on developing models but lacked user-friendly tools to facilitate the application of their findings in real-world scenarios. We developed a user-friendly web application that integrates the trained models, allowing users to generate data, predict energy production, and assess building energy performance. This provides a practical and intuitive way for stakeholders to utilize our research.

By addressing these limitations, our work aims to contribute to the development of more accurate, robust, and sustainable AI-powered tools for predicting the energy production of PV systems and supporting the transition towards a cleaner and more efficient energy future.

2.4 Conclusion

In this chapter, we conducted a literature review of the research conducted in the field of renewable energy, particularly solar energy, and the use of machine learning and deep learning techniques to predict the energy production of photovoltaic (PV) systems. We discussed the importance of accurate energy production predictions for optimizing the performance of PV systems, the challenges associated with these predictions, and the various approaches proposed in the literature to address these challenges. We explored the potential of synthetic data generation to enhance the performance of machine learning models and the impact of training these models on the environment. The research presented by Alomari et al. [18], Bowoo et al. [19], Luo et al. [20], Pamela et al. [21], and Abhishek et al. [22] has made significant contributions to the field, but there are limitations that our work aims to address. By generating a larger and more diverse dataset, incorporating BIM data, shifting from forecasting to prediction, measuring energy consumption and carbon emissions, and developing a user-friendly web application, our work aims to develop more accurate, robust, and sustainable AI-powered tools for predicting the energy production of PV systems and supporting the transition towards a cleaner and more efficient energy future.

Chapter 3

Reproducing the existing approach

In this chapter, the first step is to reimplement what Youssef & Ilyas have done in their research. Following that, the focus will shift to the main part of our work, concentrating primarily on aspects not covered by Youssef & Ilyas. The challenges faced and the solutions found to overcome them will also be discussed.

3.1 Reimplementing Yousef & Ilyas' work

Youssef & Ilyas' research focused on developing a novel approach that combines photogrammetry and deep learning techniques to address the problem of BIPV (Building-Integrated Photovoltaics) decarbonization. Their work aimed to enhance the accuracy of energy production predictions for BIPV systems by leveraging the power of AI and ML. They are calling their method BIM-AITIZATION referring to the integration of BIM data, AI techniques, and automation principles. The following sections undertake a reimplementation of their work to foster a deeper understanding of the problem, their proposed solution, and to inform the development of our own approach.

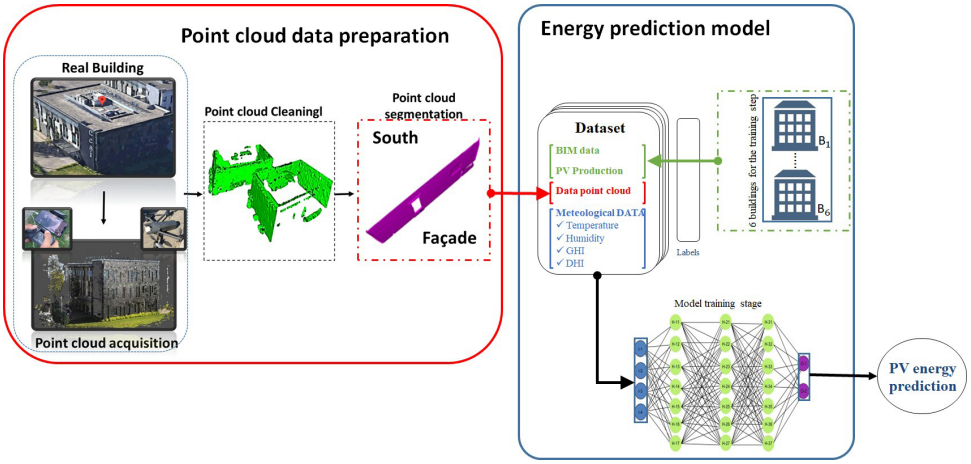

Figure 23: An overview of the BIM-AITIZATION approach [5].

3.1.1 Data preparation

To develop their model, Youssef & Ilyas created a dataset that contains the real energy production for several building located at Strasbourg. The dataset contains different types of variables, we have meteorological data, and BIM data. The meteorological data includes temperature, humidity, wind speed, and solar radiation. The BIM data includes the facade's area, the window's area, the latitude and longitude of the building.

The dataset was assembled manually from different sources and contains the real energy production for each building. The goal is to predict the energy production using the meteorological and BIM data.

3.1.1.1 Meteorological data

The meteorological data is collected from two sources: Copernicus Atmosphere Monitoring Service (CAMS) and Open Data Soft (OPDS). The data obtained from those websites was processed to have daily values of temperature, humidity, wind speed, and solar radiation, etc. The data was then merged into a single dataset.

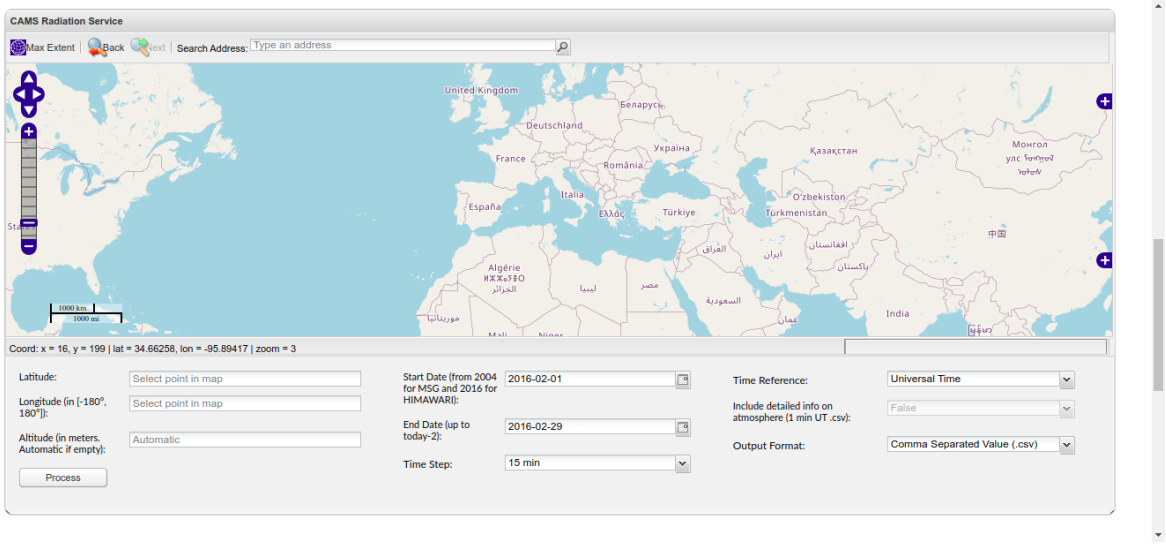

Figure 24: An overview of the CAMS website [6].

Figure 24 provides an overview of the CAMS website, one of the sources used to obtain the required meteorological data. Users can specify the location, desired date range, time step (from 15 minutes to monthly), and output format. Downloaded data is then processed to obtain daily values for relevant variables.

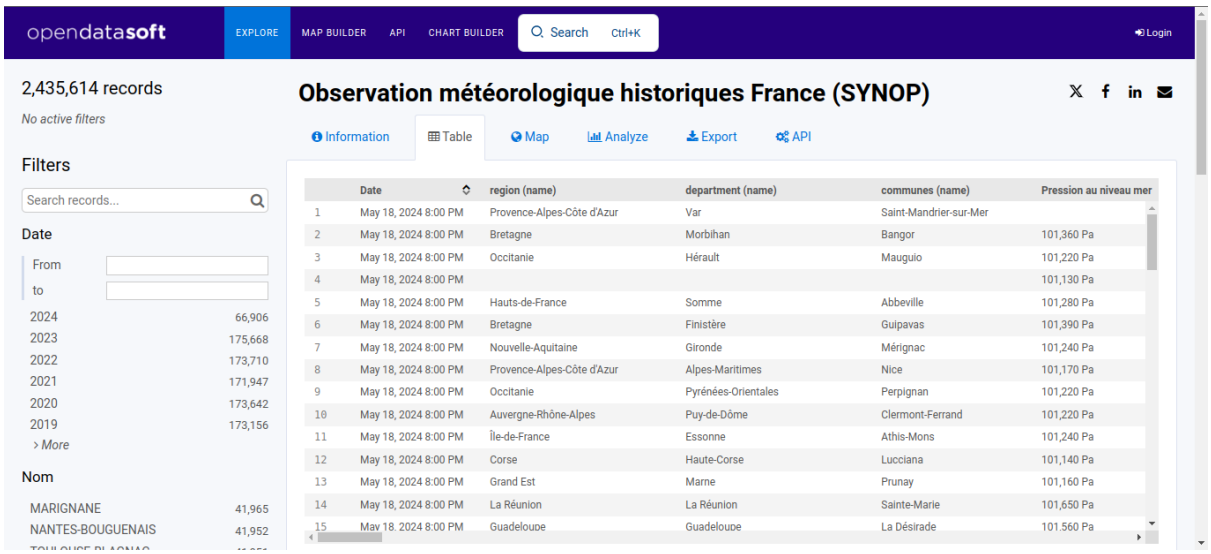

Figure 25: An overview of the OpenDataSoft website [7].

Similarly, Figure 25 depicts the OpenDataSoft website, which provided supplementary meteorological variables to complete the dataset.

3.1.1.2 BIM data

To collect the BIM data, Youssef & Ilyas used IFC (Industry Foundation Classes) files. The IFC files contain the building's geometry, the facade's area, the window's area, the latitude, and the longitude of the building. The data was then processed to extract the necessary information and merged into the dataset.

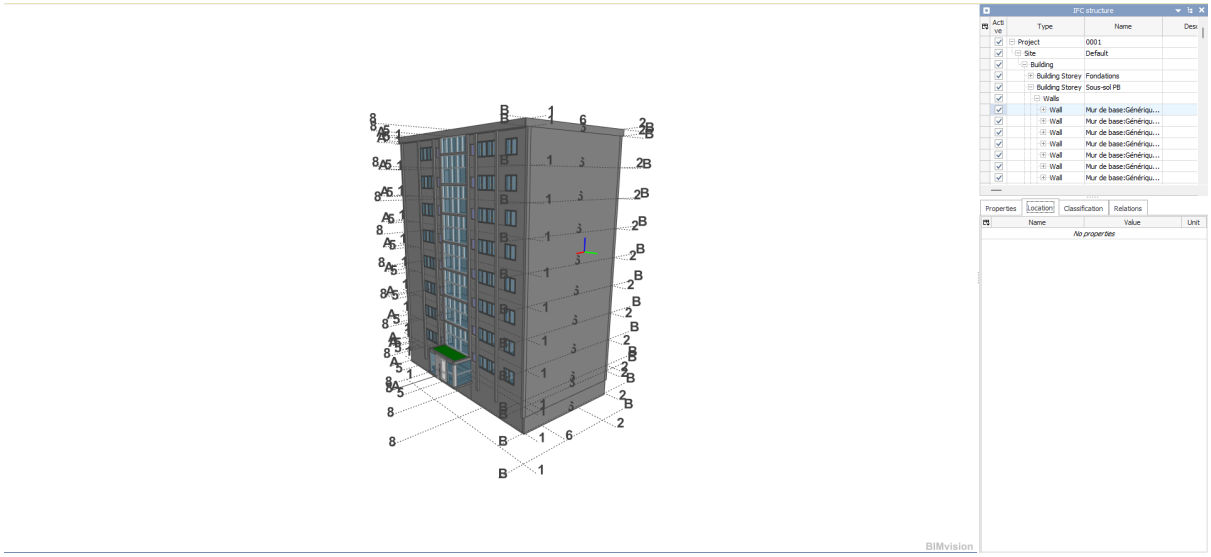

Figure 26: An example of an IFC file in BIM Vision.

Figure 26 illustrates an example of an IFC file viewed in BIM Vision, a software designed for visualizing these files. The richness of information within an IFC file is evident, encompassing details about the entire building structure.

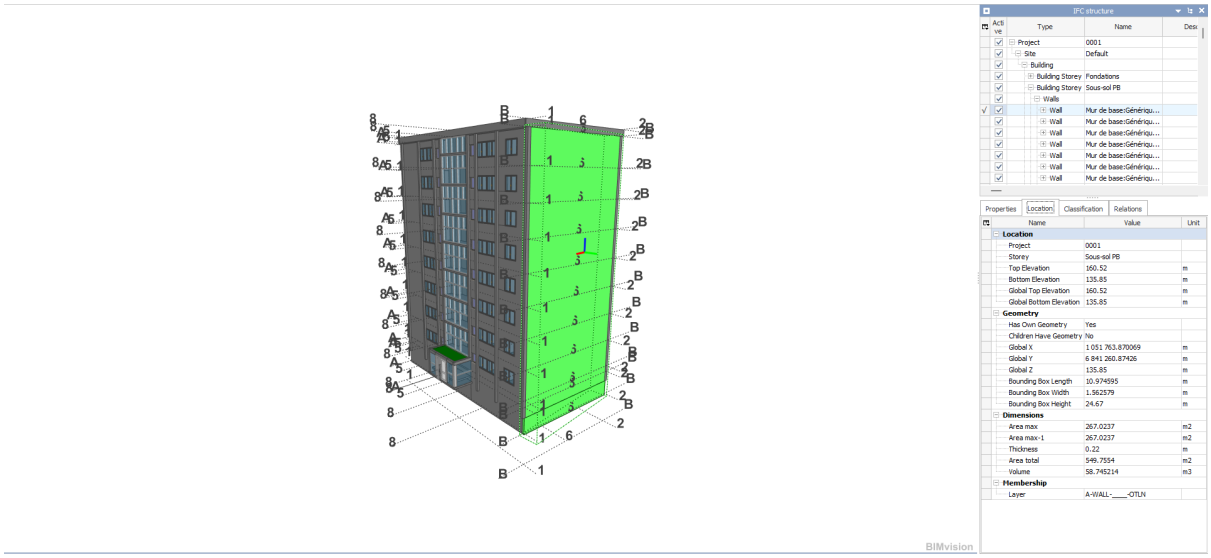

Figure 27: An example of a selection in BIM Vision.

Extracting specific data from IFC files is facilitated by BIM Vision's selection tools. Figure 27 showcases this process; selecting a facade highlights it in green, allowing for the extraction of relevant information about the facade from the right panel.

3.1.2 The full dataset

After talking about the types of data, we will now present the full dataset that Youssef & Ilyas used in their research. The dataset contains the following variables:

Table 1: The description of the parameters in the dataset.

Parameter	Description	Unit
CAMS_TOA	Irradiation on horizontal plane at the top of atmosphere	Wh/m^2
CAMS_clearskyGHI	Clear sky global irradiation on horizontal plane	Wh/m^2
CAMS_clearskyBHI	Clear sky beam irradiation on horizontal plane	Wh/m^2
CAMS_clearskyDHI	Clear sky diffuse irradiation on horizontal plane	Wh/m^2
CAMS_clearskyBNI	Clear sky beam irradiation on mobile plane	Wh/m^2
OPDS_PresNmer	Pressure	mbar
OPDS_VariaPres3h	Pressure variation in 3 hours	mbar
OPDS_TendanceBaro	Barometric trend	—
OPDS_Dir-vent	Wind direction	Degree
OPDS_Vit-vent	Wind speed	m/s
OPDS_Temp	Extreme air temperature	$°C$
OPDS_Pr-Rosekelvin	Dew point	Kelvin
OPDS_Humidity	Relative humidity	%
OPDS_Vis-Horiz	Horizontal visibility	m
OPDS_Temps-Present	Present time	—
OPDS_NebulositeNuagesInf	Nebulosity of lower level clouds	OCTA
OPDS_VariaPrese24h	Pressure variation in 24 hours	mbar
OPDS_Rafale	Gusts of wind over a period	m/s
OPDS_Precip24h	Precipitable water	cm
BIM_Latitude	Latitude, geographical coordinates	DMS
BIM_Longitude	Longitude, geographical coordinates	DMS
BIM_SurfaceSUD Facade	Surface of the south facade	m^2
Tecsol-Ensoleillement	Sunshine irradiance measured	kWh/m^2
Tecsol-Production	PV power generation	kWh

To ensure that the dataset is ready for training, Youssef & Ilyas performed some pre-processing steps, such as handling missing values, normalizing the data, and splitting it

into training and testing sets. These steps are crucial to ensure the model's accuracy and generalization on unseen data. These steps are detailed in the next section.

3.1.3 Data processing

3.1.3.1 Handling missing values

One of the first steps in data preprocessing is handling missing values. Missing values can affect the performance of machine learning models. To deal with missing values, generally the following approaches are used:

- **Removing missing values:** In this approach, rows or columns with missing values are removed from the dataset. This approach is simple but can lead to a loss of valuable information.
- **Imputing missing values:** In this approach, missing values are replaced with a specific value, such as the mean, median, or mode of the column. This approach is more robust and can help retain valuable information.
- **Using predictive models:** In this approach, missing values are predicted using other features in the dataset. This approach is more complex but can lead to better results.
- **Using domain knowledge:** In some cases, domain knowledge can be used to fill in missing values. For example, if a temperature reading is missing, it can be filled in based on the time of day or season.

3.1.3.2 Normalizing the data

Normalizing the data is an essential step in data preprocessing. Normalization ensures that all features have the same scale, which can help improve the performance of machine learning and deep learning models. There are several methods for normalizing data, including:

- **Min-Max scaling:** This method scales the data to a fixed range, usually between 0 and 1. It is calculated as follows:

$$X_{\text{norm}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}} \tag{1}$$

- **Standardization:** This method scales the data to have a mean of 0 and a standard deviation of 1. It is calculated as follows:

$$X_{\text{norm}} = \frac{X - \mu}{\sigma} \tag{2}$$

  μ is the mean of the feature we are trying to scale and σ is the standard deviation.

- **Scaling to unit length:** A common approach in machine learning to normalize a feature vector to have a unit length is vector normalization. This technique scales the components of the feature vector so that the overall vector has a magnitude of one. It involves dividing each component of the vector by the Euclidean norm (length) of the vector, as shown below:

$$X_{\text{norm}} = \frac{X}{\|X\|} \tag{3}$$

3.1.3.3 Splitting the data

To train and evaluate machine learning models, the dataset is typically split into training and testing sets. The training set is used to train the model, while the testing set is used to evaluate its performance. The data is usually split in a ratio of 70% training and 30% testing, but this can vary depending on the size of the dataset and the complexity of the model.

Because Yousef & Ilyas' dataset contains data for 7 buildings, they took 6 buildings for training and 1 building for testing. This approach ensures that the model is trained on a diverse set of data and can generalize well to unseen buildings.

3.1.4 Conclusion

This section provided a comprehensive overview of the dataset employed by Youssef & Ilyas in their research, along with the preprocessing steps implemented to prepare the data for training. With a firm grasp of the data and preprocessing techniques, we can now transition to the subsequent section, which delves into the modeling aspect.

3.2 Modeling

In this section, the modeling part of Youssef & Ilyas' research will be discussed. The difference between forecasting and prediction will be presented first. Next, the different types of models used in their research will be examined. Finally, the evaluation metrics used to assess the performance of the models will be discussed.

3.2.1 Introduction

In the modelling phase, we have 2 different problems: forecasting, and predicting. For forecasting, we will take the energy production for one building over one year or more, train a model to learn the following features (seasonality, trend, or a mix of both) in that window of data. The trained model will be used to generate a forecast for 1 to 7 days into the future. We can't forecast for more than that because the error will start accumulating.

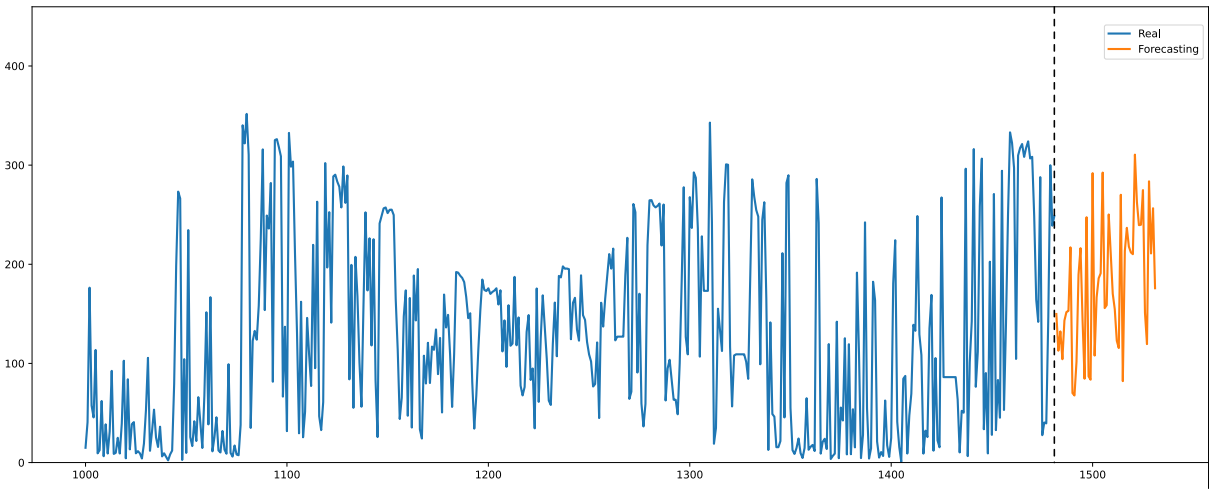

Figure 28: An example showcasing the use of forecasting.

On the other hand, the prediction problem will not use any historical data to train the model. Instead, the data will be shuffled, and only the features for a given day will be used to predict the energy produced on that day. Rather than having one model for each building, a single model will be used to learn how to associate the input features with the output.

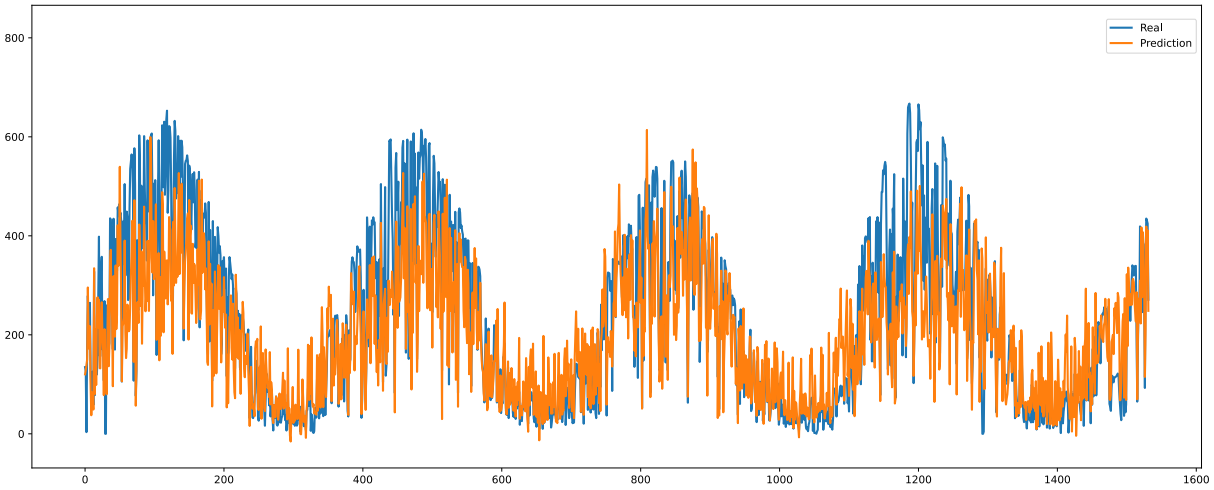

Figure 29: An example showcasing the prediction problem.

Youssef and Ilyas worked on the forecasting problem, this is why in this modeling section in this chapter we will talk about only the forecasting problem.

3.2.2 Structuring the data

Training a model to give accurate forecasts involves structuring the data in a way that will preserve important characteristics such as seasonality and trend in the data. This can be accomplished by splitting the data into windows of a defined size, where the input consists of a sequence of consecutive time periods, and the output is the subsequent time period. For instance, if the window size is set to 10, the input would contain data from 10 consecutive days, and the output would be the data for the 11th day.

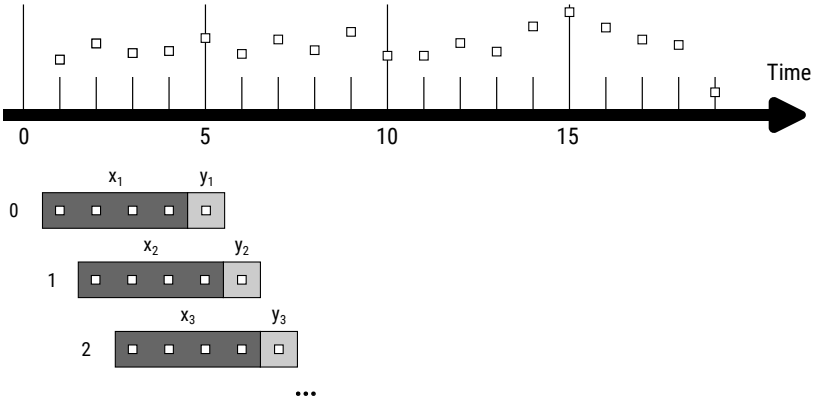

Figure 30: Splitting the data into a series of windows.

Figure 30 illustrates how the data is split into windows of size 4. The input data consists

of the first 4 days, and the output data is the 5th day. This process is repeated for the entire dataset, creating a series of input-output pairs that can be used to train the model.

In general, if we start with a dataset with the shape (n, m), where n is the number of samples and m is the number of features, after splitting the data into windows, we will have a new dataset with the shape $(n - w, w, m)$, where w is the window size.

3.2.3 Sequence models

Sequence models are a class of models that are designed to handle sequences of data, such as time series data. These models are well-suited for forecasting tasks, as they can capture the temporal dependencies in the data. Some common sequence models include Long Short-Term Memory (LSTM) networks, Recurrent Neural Networks (RNNs), and Gated Recurrent Units (GRUs). Their ability to learn and model patterns across sequential inputs makes them a good choice for forecasting applications.

3.2.3.1 Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are a class of neural networks that are designed to handle sequential data. RNNs have a feedback loop that allows information to persist across time steps, making them well-suited for tasks such as time series forecasting. However, RNNs can suffer from the vanishing gradient problem, which can make it difficult for the model to learn long-term dependencies in the data.

The vanishing gradient problem occurs when the gradients of the loss function with respect to the weights become very small, causing the weights to stop updating. This can prevent the model from learning long-term dependencies in the data.

The RNN network is composed of a single cell that takes an input x_t at time step t and produces an output o_t. The cell also has a hidden state h_t that represents the network's memory at time step t and acts as memory of the network.

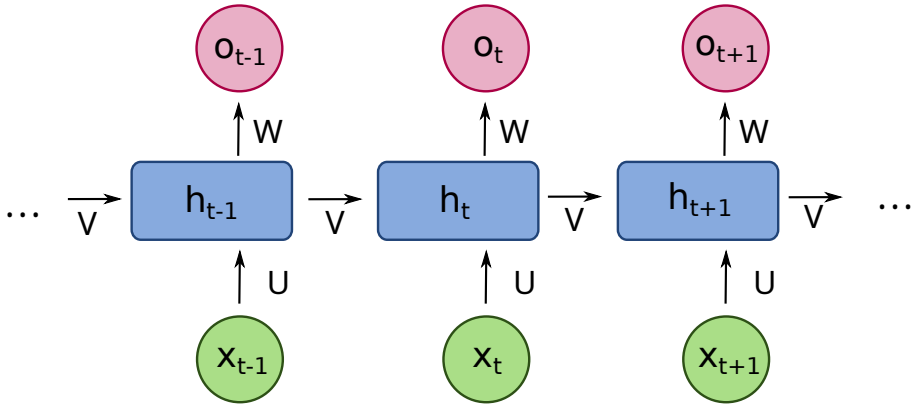

Figure 31: The architecture of an RNN cell [8].

Figure 31 shows the architecture of an RNN cell and the stacking of these cells to form a complete RNN network.

3.2.3.2 Long Short-Term Memory networks

Long Short-Term Memory networks (LSTMs) were first introduced by Hochreiter & Schmidhuber in 1997 [9]. They are a special kind of RNN, capable of learning long-term dependencies. From the beginning, LSTMs were explicitly designed to avoid the long-term dependency problem that RNN networks faced. This is why LSTMs have become popular because remembering information for long periods of time is what they excel at.

LSTMs have a more sophisticated module (unit) than RNNs, this module is composed of a cell, an input gate, an output gate, and a forget gate. The cell remembers values over arbitrary time intervals and the three gates regulate the flow of information into and out of the cell.

- **Forget gate:** Decides what information to throw away from the cell state. The value of the forget gate is between 0 and 1, where 0 means "completely forget" and 1 means "completely keep".
- **Input gate:** Decides what new information to store in the cell state.
- **Output gate:** Decides what to output based on the cell state by assigning a value between 0 and 1 to each value in the cell state.

The following figure explains the architecture of an LSTM cell. These individual cells are then stacked to form a complete LSTM network.

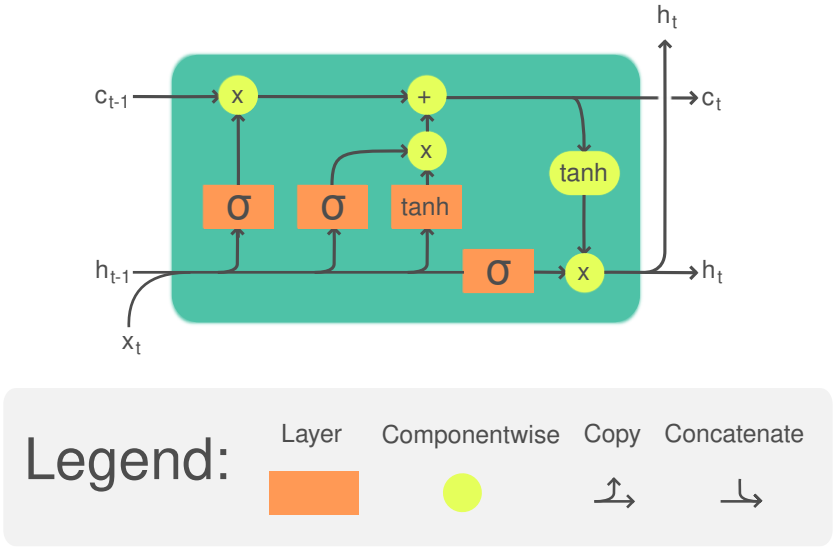

Figure 32: The architecture of an LSTM cell [10].

3.2.3.3 Gated Recurrent Units

Gated Recurrent Units (GRUs) are a type of RNN that are similar to LSTMs but have a simpler architecture. GRUs were introduced in 2014 by Kyunghyun Cho et al. [11]. GRUs have fewer parameters than LSTMs and are faster to train. They have two gates: a reset gate and an update gate. The update gate regulates the flow of data between time steps, while the reset gate determines how much previous data is retained or discarded.

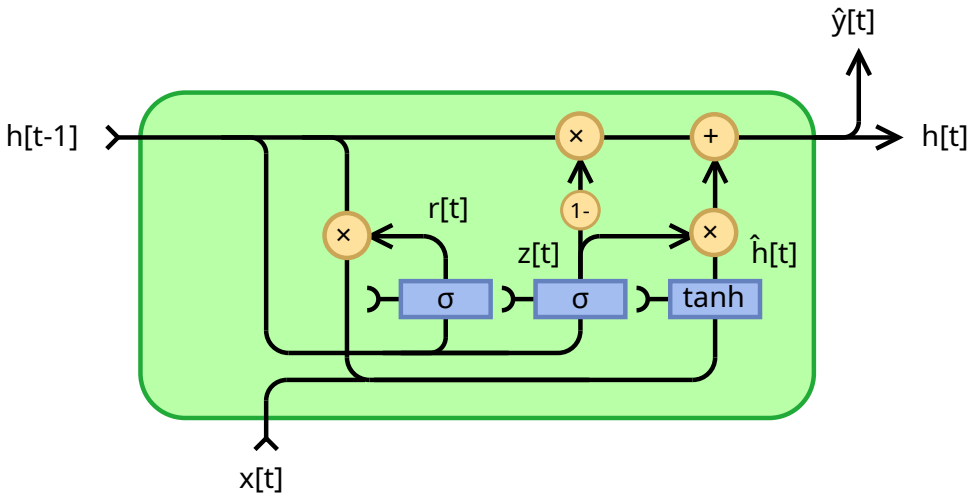

Figure 33: The architecture of a GRU cell [12].

Figure 33 shows the architecture of a GRU cell. The input at time step t is denoted as x_t, the hidden state from previous time step $t-1$ is denoted as h_{t-1}. The cell outputs the output $\hat{y}_t$ and the hidden state h_t.

The reset gate is denoted as r_t and the update gate is denoted as z_t. The cell also has a candidate hidden state $\hat{h}_t$ that is used to update the hidden state h_t.

3.3 Experiments

3.3.1 Experiments settings

As previously noted, Youssef and Ilyas initially assembled real data from 7 buildings into a single large dataset. After my arrival, we included data from a tower in Strasbourg, expanding the dataset to cover 8 buildings. The updated dataset now consists of 10,236 rows and 39 features.

	production	latitude	longitude	vmp	imp	voc	isc	p_per_m2	p_max	panel_area	...	surface_pressure	wind_direction_100m	wind_direction_10m	wind_speed
0	0.5	48.575678	7.765640	27.3	7.7	33.3	8.17	143.0	210	1.72	...	1031.470213	256.510638	245.893617	5.(
1	19.6	48.575678	7.765640	27.3	7.7	33.3	8.17	143.0	210	1.72	...	1034.381250	317.458333	308.666667	6.2
2	1.1	48.575678	7.765640	27.3	7.7	33.3	8.17	143.0	210	1.72	...	1037.814583	291.479167	256.937500	2.1
3	5.9	48.575678	7.765640	27.3	7.7	33.3	8.17	143.0	210	1.72	...	1036.620833	226.937500	256.583333	2.(
4	0.1	48.575678	7.765640	27.3	7.7	33.3	8.17	143.0	210	1.72	...	1032.268750	259.895833	249.145833	5.2
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
10231	17.4	48.575793	7.766982	27.3	7.7	33.3	8.17	143.0	210	1.72	...	1016.602083	134.250000	125.750000	1.5
10232	25.5	48.575793	7.766982	27.3	7.7	33.3	8.17	143.0	210	1.72	...	1017.581250	130.666667	114.333333	2.5
10233	22.3	48.575793	7.766982	27.3	7.7	33.3	8.17	143.0	210	1.72	...	1016.029167	184.333333	176.937500	3.2
10234	20.4	48.575793	7.766982	27.3	7.7	33.3	8.17	143.0	210	1.72	...	1018.052083	238.083333	230.645833	3.(
10235	2.0	48.575793	7.766982	27.3	7.7	33.3	8.17	143.0	210	1.72	...	1016.402083	231.312500	224.270833	3.7

10236 rows × 39 columns

Figure 34: An overview of the 8 buildings dataset.

The dataset is divided into three splits, ensuring the sequential aspect of the data is preserved rather than splitting it randomly. Using the `building_id` column, we first split the data into two parts: one for training and the other for testing, with one building reserved for the testing phase. Then, the training portion is further divided into two parts: training and validation.

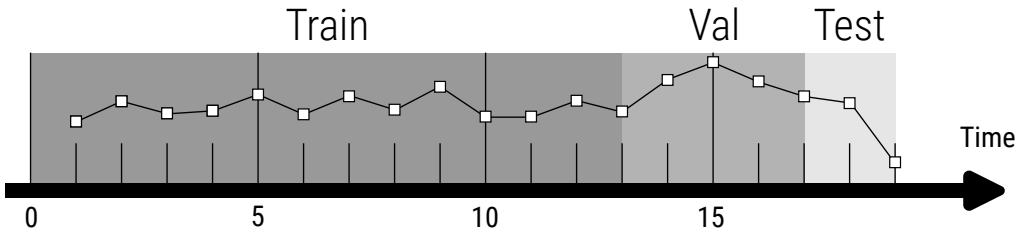

Figure 35: An overview of the data splitting process.

After splitting the data, we preprocess it by scaling the features. We then organize the data into windows of size 10, changing the input shape from $(8982, 38)$ to $(8972, 10, 39)$. The number of features increased by one because we added the energy production column. We do this to enable our model to use past values to predict the future.

After preparing the data, we can use it to train three distinct deep learning models: a Long Short-Term Memory (LSTM) model, a Convolutional Neural Network (CNN) model, and a Hybrid model, which we will refer to as ConvLSTM1D. For the LSTM model, the appropriate input shape is $(n, window\_size, m)$, where n represents the number of samples, $window\_size$ corresponds to the length of the time series sequence, and m denotes the number of features. Regarding the CNN model, the data needs to be reshaped into the format $(n, 1, m)$, with the additional dimensions representing the width or height of the 1D input data. Finally, the ConvLSTM1D model necessitates an input shape of $(n, window\_size, 1, m)$, where the additional dimension of 1 represents the width or height of the 1D input data.

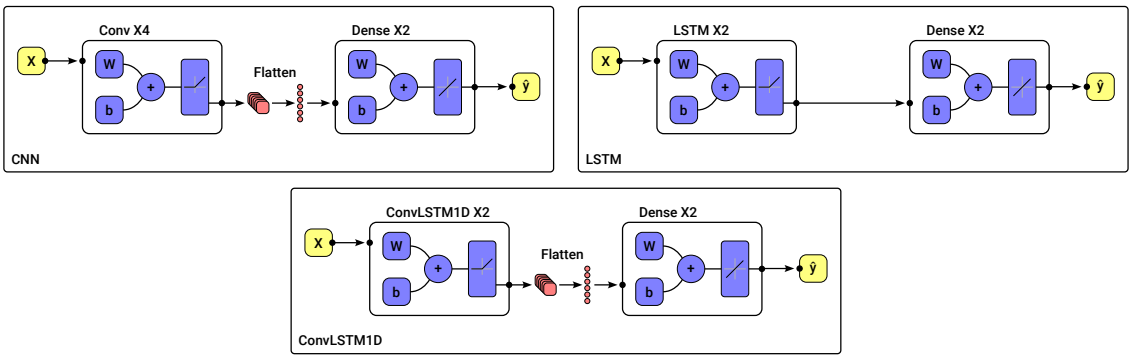

Figure 36: The architectures of the LSTM, CNN, and ConvLSTM1D models.

The architectures of the LSTM, CNN, and ConvLSTM1D models are depicted in Figure 36. At the top, we have the CNN model's architecture, followed by the LSTM model's architecture below it. Finally, at the bottom, we have the architecture of the ConvLSTM1D model, which is a hybrid combination of the CNN and LSTM models.

During the training phase, we use the Adam optimizer, with the learning rate set to 0.001. The loss we are trying to optimize during training is the MAE. We train the models for 300 epochs max, with a batch size of 128. The models are evaluated using the MAE, the Mean Squared Error (MSE) metrics, the Mean Absolute Percentage Error (MAPE), and the R-squared score.

Alongside the deep learning models, we also trained some machine learning models such as Random Forest, Decision Tree, Linear Regression, Gradient Boosting, and XGBoost. The results of the experiments are presented in the next section.

3.3.2 Results

In this section, the results of the experiments will be presented. The performance of the deep learning models (LSTM, CNN, and ConvLSTM1D) will be compared with the machine learning models (Random Forest, Decision Tree, Linear Regression, Gradient Boosting, and XGBoost). The models will be evaluated using the MAE, the MSE, the MAPE, and the R-squared score.

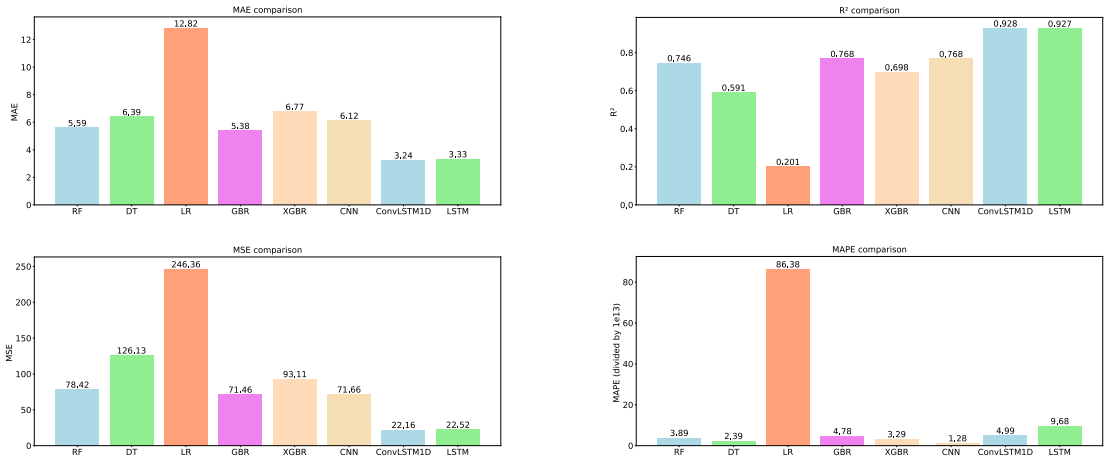

Figure 37: The results of the experiments.

From the plots in figure 37 we see that the `ConvLSTM1D` and `LSTM` models show superior performance across all metrics. Specifically, `ConvLSTM1D` achieves the lowest MAE (3.24) and MSE (22.16), alongside the highest R^2 value (0.928) and a MAPE equal to 1.28. `LSTM` follows closely with similar performance, showing slightly higher MAE (3.33) and MSE (22.52), a comparable R^2 value (0.927), and a MAPE of 9.68. These results indicate that `ConvLSTM1D` and LSTM are highly effective for this particular prediction task, offering high accuracy and low error rates.

In contrast, `Linear Regression` (LR) performs the worst among all models. It has the highest MAE (12.82) and MSE (246.36), the lowest R^2 value (0.201), and the highest MAPE (86.38). This poor performance across all metrics suggests that `LR` is not suitable for the given dataset and task. The significant errors and low R^2 value imply that `LR` fails to capture the underlying patterns and relationships in the data.

The remaining models—`RF`, `DT`, `GBR`, `XGBR`, and `CNN`—demonstrate intermediate performance. Among these, `GBR` and `CNN` show relatively better performance, with moderate MAE, MSE, and R^2 values. `RF` and `XGBR` are slightly behind, but still perform reasonably well compared to `LR`. For example, `RF` achieves an MAE of 5.59 and an R^2 value of 0.746, while `GBR` and `CNN` both have an R^2 value of 0.768.

Just looking at the numbers we showed in figure 37 isn't enough to judge the performance of the models. We need to show the predictions on the test set, this way just by observing the plots we can spot which model is performing well.

In figure 38, we are comparing between `Random Forest` and `ConvLSTM1D` models. We are testing both models on the test set, this set wasn't used in training, this is a good test to see if the models can generalize to unseen data or not. At first glance, it seems like that both models are following the data well, but when looking closely we see that sometimes the `Random Forest` model is failing to give good predictions. On the other hand, the `ConvLSTM1D` model is performing way better than `Random Forest`, and this is what the performance metrics indicated.

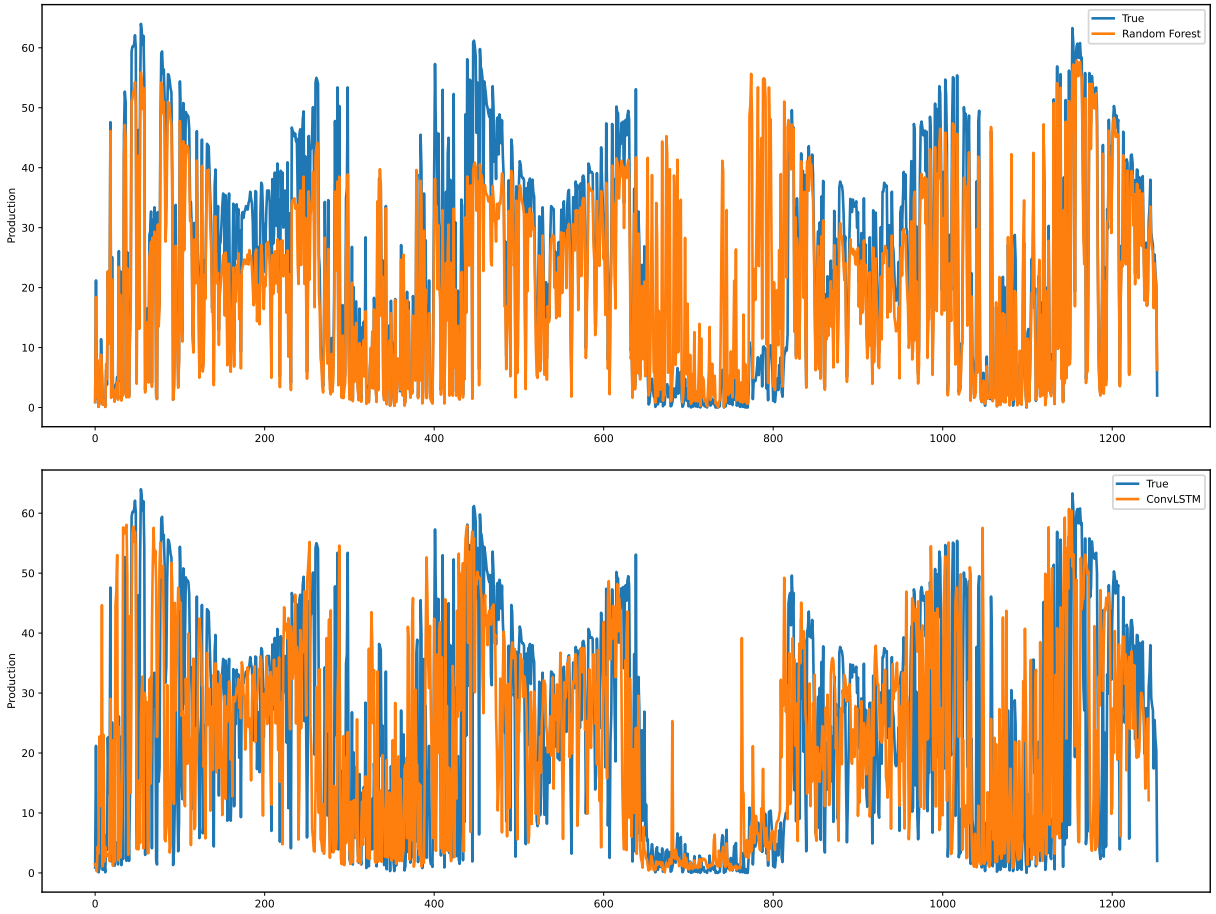

Figure 38: A comparison between `Random Forest` and `ConvLSTM1D` on the test set.

Let's also look at a study case of the CESI building, an image of the building is shown in figure 39. We don't have the real data for the energy production, but we used a software called BimSolar [13] to simulate the energy production for the building.

Figure 39: The CESI building.

We imported the CESI building into the software (see figure 40), added the solar panels to the south facade, and then we simulated the energy production as shown in figure 40.

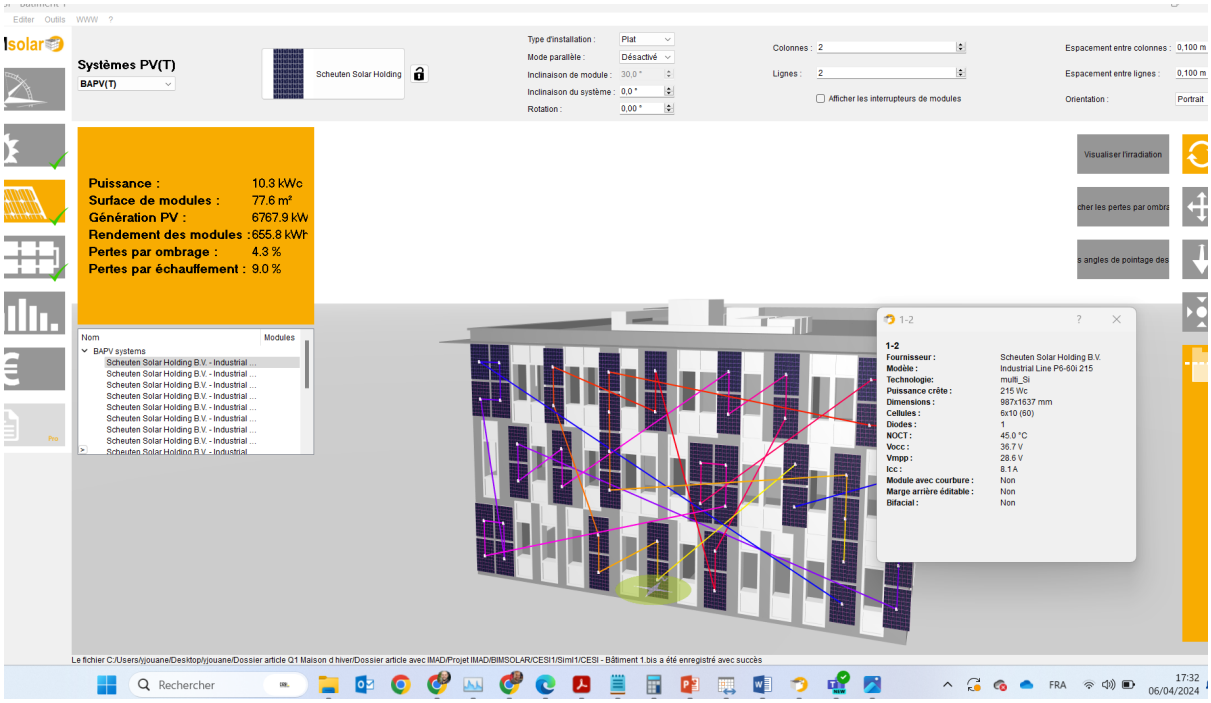

Figure 40: The CESI building in BimSolar.

The energy production of the CESI building in BimSolar can be set to multiple frequencies, such as daily or monthly. We set the frequency to daily and obtained the energy production for the building, as shown in the following figure.

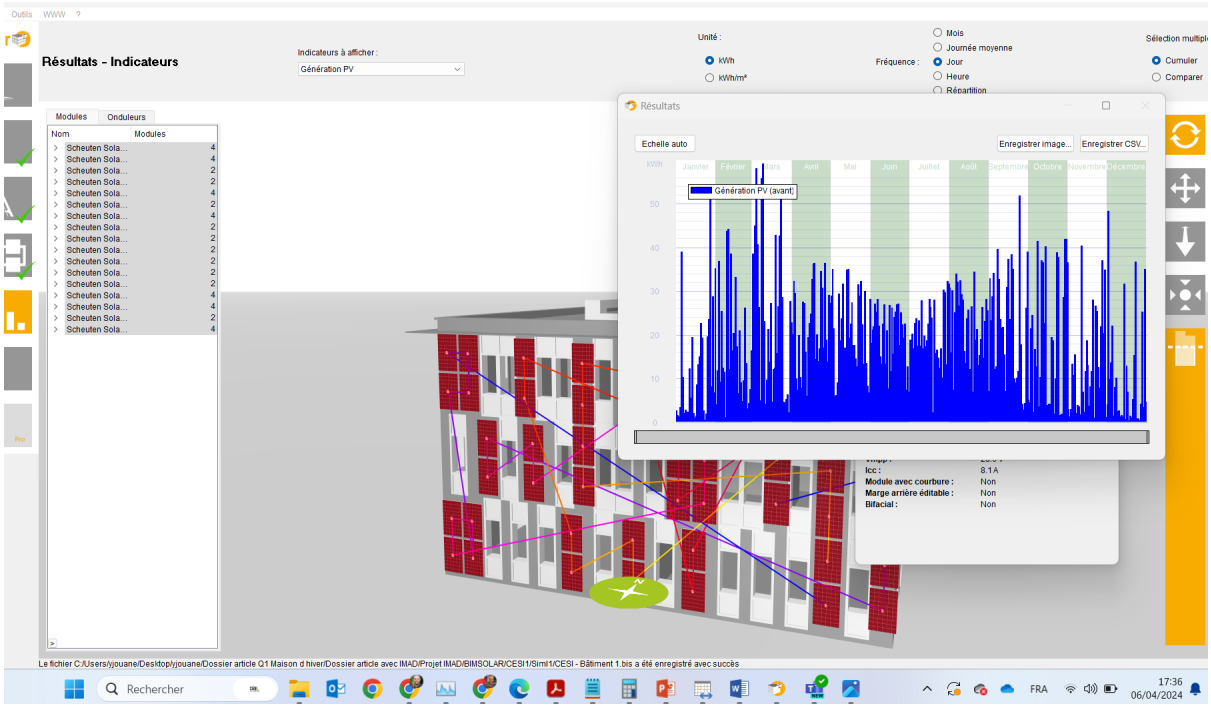

Figure 41: The energy production of the CESI building in BimSolar.

After obtaining the energy production data, we used one of our models to predict the energy production for the CESI building. The predictions are shown in the following figure.

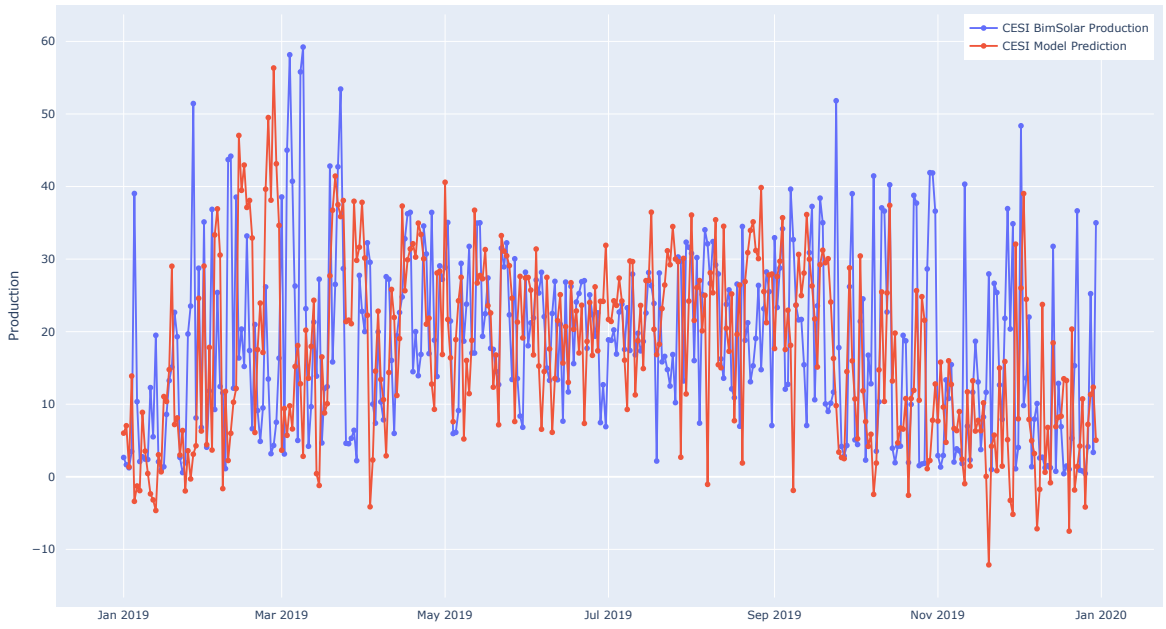

Figure 42: The energy production of the CESI building compared to the model's predictions.

From the figure, we can see that the model's predictions are close to the simulated energy production of the CESI building. This indicates that the model is performing well and

can accurately predict the energy production of buildings with solar panels installed on the south facade and that are located in Strasbourg.

3.3.3 Limitations

The results obtained in the previous section, while promising, do not necessarily imply that the trained models or the dataset used for training are perfect. For instance, if a new solar panel type is introduced, the model will consistently provide the same prediction because the dataset lacks features related to solar panels. Another limitation of Youssef & Ilyas' solution is that if the model is applied outside of Strasbourg, its performance will be suboptimal due to the absence of data for buildings outside Strasbourg. Furthermore, since the model was trained on buildings with solar panels installed on the south facade, if the solar panels are installed on the roof (where the energy production is generally higher than the facade), the model will not adjust to this change. These limitations highlight the shortcomings of Youssef & Ilyas' approach, and the subsequent sections will aim to address these issues.

Let's illustrate the limitations of the model with an example. We have a building in Toulouse, France, with solar panels installed on the roof. We don't have access to real data but we did simulate the energy production for the building using BimSolar. Then we tried to compare the model to the BimSolar simulation, the results are shown in the following figure.

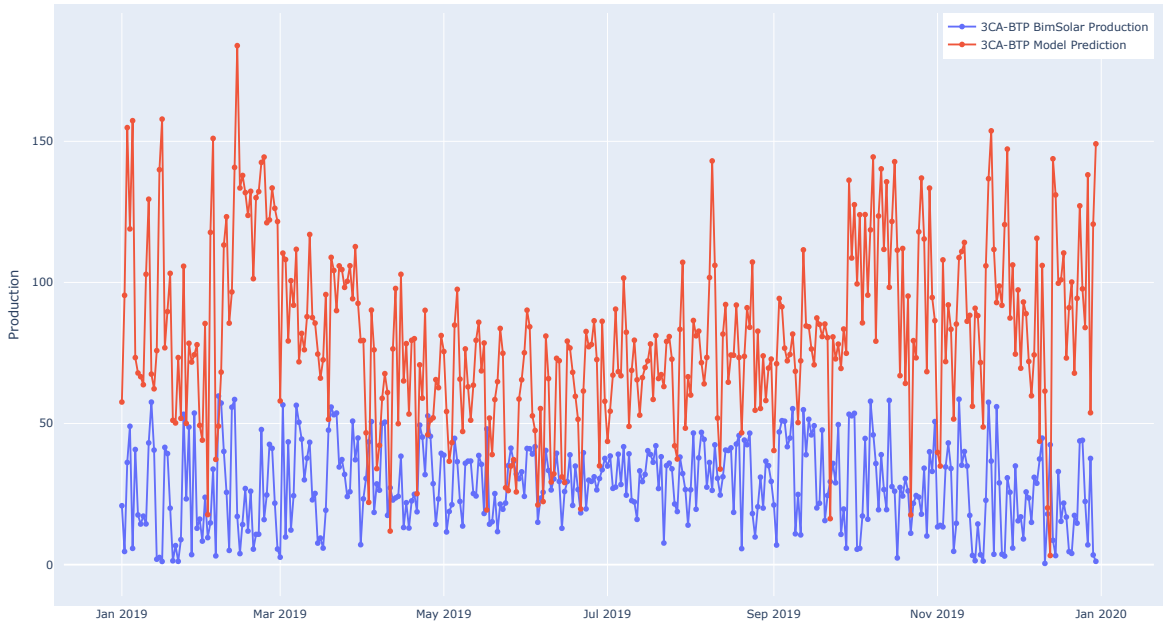

Figure 43: The energy production of the Toulouse building compared to the model's predictions.

From the figure, we can see that the model's predictions are not close to the actual energy production of the Toulouse building. This indicates that the model is not performing well and cannot accurately predict the energy production of buildings with solar panels installed on the roof. This example highlights the limitations of the model and the need for improvements to address these issues.

3.4 Conclusion

Building upon Youssef & Ilyas' work, we reimplemented their approach and highlighted the distinction between forecasting and prediction. While their models achieved promising results, further improvements are possible. Integrating solar panel data, expanding the dataset, and focusing on prediction over forecasting offer significant potential. Shifting to a prediction problem offers significant advantages, as demonstrated by BimSolar's simulation capabilities. This allows for direct energy production estimation without historical data, a valuable feature for replacing sizing softwares with machine learning or deep learning models. These advancements will be explored in the next section.

Chapter 4

Our novel approach

This chapter delves into the newly curated dataset, exploring the changes made, additions included, and the crucial aspect of diversification. Following this, we move to the modeling phase, specifically focusing on tackling the prediction problem. Finally, we present the obtained results and conclude the chapter.

4.1 Synthetic data

To address the limited data available in Youssef & Ilyas' work, we explored generating synthetic data using BimSolar. Since BimSolar's estimations closely resemble reality, we aimed to leverage it to train our model. We achieved this by creating a massive dataset where we varied building types, the number and types of solar panels, their placement on the roof or the south facade, and the building's location. This process resulted in a dataset ten times larger than the one used by Youssef & Ilyas.

Furthermore, we introduce a novel approach to data collection, focusing on automation. This significant addition simplified the process of gathering data from diverse sources, significantly accelerating the dataset assembly and contributing to its substantial size.

4.1.1 Meteorological data

To achieve the automation of meteorological data access, we explored APIs that allow programmatic data retrieval. Two promising solutions emerged: NASA POWER and the Solcast project.

4.1.1.1 NASA Power

NASA POWER offers solar and meteorological datasets derived from NASA research, catering to the needs of renewable energy, sustainable buildings, and agriculture. The meteorological data originates from two key sources:

- **NASA's GMAO MERRA-2 assimilation model:** This provides long-term historical data, spanning from 1981 to near real-time (approximately one month behind). MERRA-2 is a version of NASA's Goddard Earth Observing System (GEOS) Data Assimilation System.
- **GMAO Forward Processing – Instrument Teams (FP-IT) GEOS 5.12.4:** This source delivers more recent data, covering the period from the end of the MERRA-2 data stream to within a few days of real-time. The POWER Project team processes this data daily and appends it to the MERRA-2 time series, ensuring low latency and availability within roughly two days of real-time.

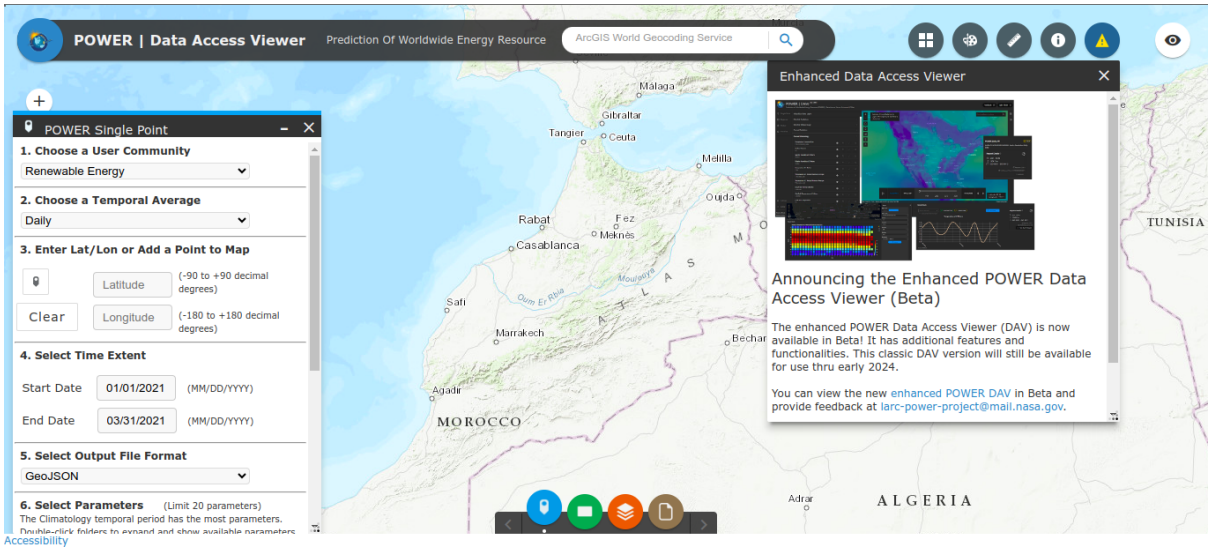

Figure 44: An overview of the NASA POWER project [14].

While the NASA POWER API offers access to a vast array of over 100 meteorological variables, including temperature, precipitation, irradiance, and pressure, it does have limitations. Notably, users can only request a maximum of 20 parameters per query. Additionally, the retrieved data may contain missing values, requiring further processing.

Figure 45 shows the steps we need to follow to access meteorological data through the NASA Power API with Python:

1. **Define the variable names:** The names of the variables we want to retrieve are described in the NASA Power documentation.

2. **Select 20 variables:** Choose a maximum of 20 variables from your list and join them with commas to create a comma-separated string.

3. **Construct the API endpoint:** Use the base URL shown in the image and replace the placeholders with your chosen variables, latitude, longitude, start and end dates, and desired format.

```python
parameters = [
    'CLRSKY_SFC_SW_DWN',
    'ALLSKY_SFC_SW_DWN',
    ...
]

parameters_1 = ','.join(parameters[:20])
url_1 = f'https://power.larc.nasa.gov/api/temporal/daily/point?parameters={parameters_1}&community=SB&longitude=7.7700&latitude=48.5700&start=20190101&end=20230101&format=CSV'
response_1 = requests.get(url_1)
```

Python

+ Code + Markdown

```python
data_1 = response_1.text
print(data_1)
```

Python

```
-BEGIN HEADER-
NASA/POWER CERES/MERRA2 Native Resolution Daily Data
Dates (month/day/year): 01/01/2019 through 01/01/2023
Location: Latitude  48.57   Longitude 7.77
Elevation from MERRA-2: Average for 0.5 x 0.625 degree lat/lon region = 267.86 meters
The value for missing source data that cannot be computed or is outside of the sources availability range: -999
Parameter(s):
CLRSKY_SFC_SW_DWN       CERES SYN1deg Clear Sky Surface Shortwave Downward Irradiance (W/m^2)
ALLSKY_SFC_SW_DWN       CERES SYN1deg All Sky Surface Shortwave Downward Irradiance (W/m^2)
```

Figure 45: An example of a request to the NASA POWER API.

The API response consists of two distinct parts:

- **Header:** This section provides valuable information about the retrieved variables. It includes the variable names, detailed descriptions, and their corresponding units.
- **Data:** Following the header, we find the actual data values in the format we specified during the request (e.g., CSV).

```
-BEGIN HEADER-
NASA/POWER CERES/MERRA2 Native Resolution Daily Data
Dates (month/day/year): 01/01/2019 through 01/01/2023
Location: Latitude  48.57   Longitude 7.77
Elevation from MERRA-2: Average for 0.5 x 0.625 degree lat/lon region = 267.86 meters
The value for missing source data that cannot be computed or is outside of the sources availability range: -999
Parameter(s):
CLRSKY_SFC_SW_DWN       CERES SYN1deg Clear Sky Surface Shortwave Downward Irradiance (W/m^2)
ALLSKY_SFC_SW_DWN       CERES SYN1deg All Sky Surface Shortwave Downward Irradiance (W/m^2)
ALLSKY_KT               CERES SYN1deg All Sky Insolation Clearness Index (dimensionless)
WS2M                    MERRA-2 Wind Speed at 2 Meters (m/s)
WS10M                   MERRA-2 Wind Speed at 10 Meters (m/s)
WS50M                   MERRA-2 Wind Speed at 50 Meters (m/s)
T2M                     MERRA-2 Temperature at 2 Meters (C)
T10M                    MERRA-2 Temperature at 10 Meters (C)
TS                      MERRA-2 Earth Skin Temperature (C)
QV2M                    MERRA-2 Specific Humidity at 2 Meters (g/kg)
RH2M                    MERRA-2 Relative Humidity at 2 Meters (%)
PRECTOTCORR             MERRA-2 Precipitation Corrected (mm/day)
PS                      MERRA-2 Surface Pressure (kPa)
WD2M                    MERRA-2 Wind Direction at 2 Meters (Degrees)
WD10M                   MERRA-2 Wind Direction at 10 Meters (Degrees)
WD50M                   MERRA-2 Wind Direction at 50 Meters (Degrees)
ALLSKY_SFC_UV_INDEX     CERES SYN1deg All Sky Surface UV Index (dimensionless)
ALLSKY_SRF_ALB          CERES SYN1deg All Sky Surface Albedo (dimensionless)
ALLSKY_SFC_SW_UP        CERES SYN1deg All Sky Surface Shortwave Upward Irradiance (W/m^2)
ALLSKY_SFC_SW_DNI       CERES SYN1deg All Sky Surface Shortwave Downward Direct Normal Irradiance (W/m^2)
-END HEADER-
YEAR,MO,DY,CLRSKY_SFC_SW_DWN,ALLSKY_SFC_SW_DWN,ALLSKY_KT,WS2M,WS10M,WS50M,T2M,T10M,TS,QV2M,RH2M,PRECTOTCORR,PS,WD2M,WD10M,WD50M,ALL
2019 1 1 61 63 27 41 0 27 2 05 3 38 5 38 2 27 2 59 1 81 4 46 98 0 0 8 99 83 265 5 265 81 267 81 0 05 0 16 4 23 56 87
```

Figure 46: An example of the data retrieved from the NASA POWER API.

4.1.1.2 Solcast

Solcast is a provider of data and tools for the global solar power industry [23]. Founded in 2016, the company's core offering is high-quality, widely accessible, and forecasting data. This data, derived from a global network of weather satellites, is independently validated and available for free trial. Users can access the data within minutes through the Solcast API Toolkit.

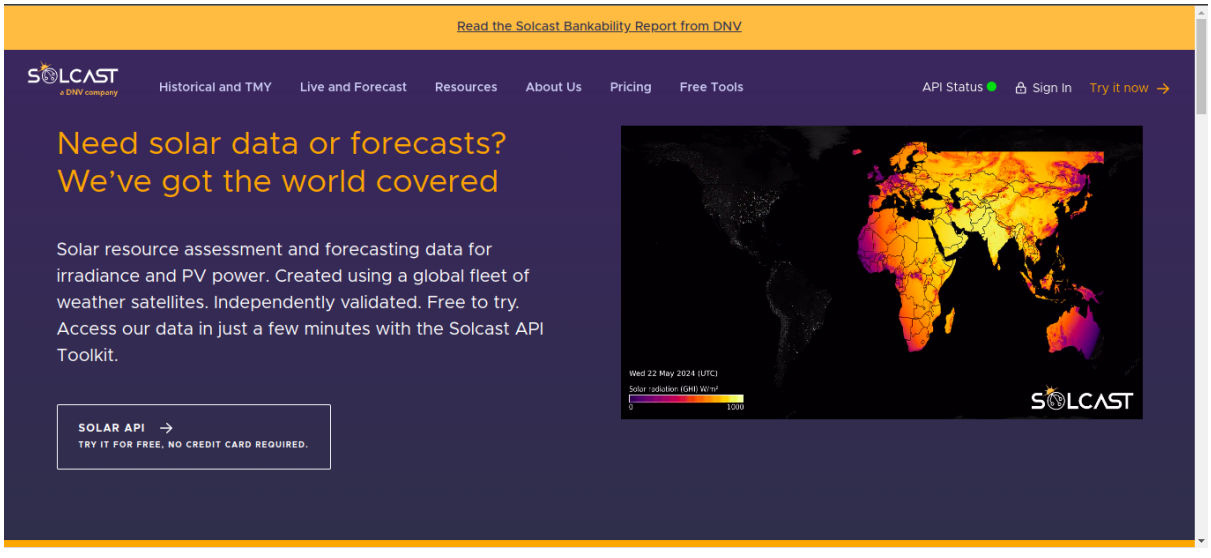

Figure 47: The solcast website [23].

The company leverages cutting-edge technology, including weather satellite imagery, machine learning, computer vision, and big data, to deliver its innovative solar forecasting and modelling solutions. Solcast processes over 600 million new forecasts every hour in a cloud-based environment, providing real-time access through an API.

Solcast's historical data offers high resolution and low uncertainty, making it ideal for our purposes. The data is readily available for integration via API and includes:

- Low uncertainty, and zero bias
- Independent validation and global coverage
- High resolution 2km data at 5-minute resolution
- PV modeling software integration (PVSyst, SAM, TMY3, CSV)
- Solar irradiance (GHI, DNI, DHI)
- Weather (Temp, Wind, Humidity, Snow, etc.)

Another benefit of using Solcast is the 6 months free access to the API given to researchers. This allowed us to access the data and compare it to the data we obtained from NASA POWER.

To access the data from Solcast, we follow the 3 steps described below:

1. **Get an API Key:** First things first, you need an API key. Just head over to Solcast, create an account, and then in your account, you'll see your unique key.
2. **Set the parameters:** Now, you need to tell Solcast what you want. You can choose the format for the data, how long you want the data for (up to 31 days at a time), and how often you want the data points (every 5, 10, 15, 20, 30, or 60 minutes). You can also set the location by specifying the latitude and longitude.
3. **Make the Request:** With your API key and parameters in hand, you can send the request to Solcast and get your data.

Let's see how to perform those steps in python. First, we provide the API Key in the headers of the request, then we set the parameters of the request, and finally, we make the request to Solcast to get the data. The flow is described in the following figure.

```python
import requests

API_KEY = '4iMv4la1yPwckcBVHI5pI8qu0Vqje5jh'

headers = {
    "Authorization": f"Bearer {API_KEY}"
}

duration = "P31D"  # maximum duration is 31 days
period = "PT60M"
format_ = "json"
time_zone = "utc"
output_parameters = "air_temp,albedo,azimuth,clearsky_dhi,clearsky_dni,clearsky_ghi,clearsky_gti,cloud_opacity,dewpoint_temp,dhi,dni,ghi,gti,p


responses = []
for date in date_range:
    start = f"{date.strftime('%Y-%m-%d')}T00:00:00.000Z"
    url = f"https://api.solcast.com.au/data/historic/radiation_and_weather?latitude={latitude}&longitude={longitude}&start={start}&duration={d

    response = requests.get(url, headers=headers)
    responses.append(response)
```

Figure 48: Making a request to the Solcast API using Python.

4.1.2 Photovoltaic data

Photovoltaic (PV) data is essential for predicting energy production from solar panels. This section outlines the solar panel characteristics that will be incorporated as new features in our dataset. It also describes the sources of this data and introduces the specific solar panels used in our training set.

4.1.2.1 Solar panel characteristics

To accurately predict energy production, we must consider the characteristics of the solar panels. Figure 49 shows 3 panels with the same type of parameters but with different characteristics.

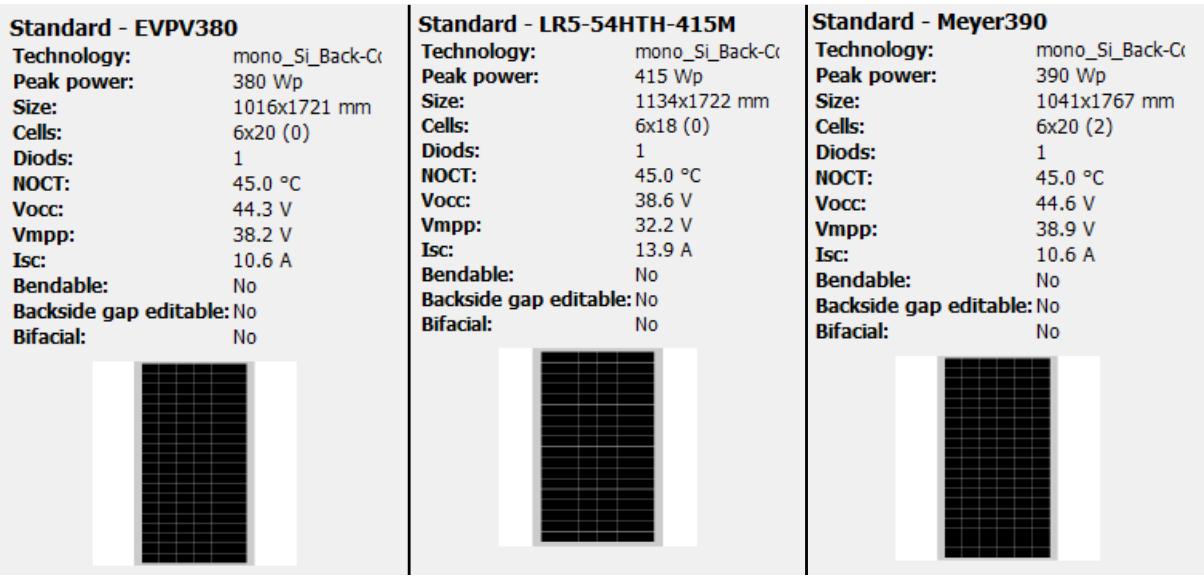

Figure 49: Three solar panels with different characteristics.

Let's understand each one of the parameters in the datashet, let's take the first panel as an example:

- **Technology:** The technology used in the solar panel, it can be monocrystalline, polycrystalline, or thin-film.
- **Peak power:** The maximum power the solar panel can produce under Standard Test Conditions (STC), which are 1000 W/m^2, 25° C cell temperature, and 1.5 air mass.
- **Size:** The dimensions of the solar panel in millimeters.
- **Cells:** This indicates the number and arrangement of solar cells within the panel. In our example, this panel has 6 rows and 20 columns of solar cells, for a total of 120 cells.
- **NOCT:** The Nominal Operating Cell Temperature, which is the estimated temperature of the solar cells under normal operating conditions, which can be significantly higher than the ambient air temperature. It is typically around 45° C.

- **Vocc:** The open-circuit voltage of the solar panel, which is the maximum voltage the panel can produce when not connected to a load.
- **Vmpp:** The voltage at which the solar panel produces the maximum power.
- **Isc:** The short-circuit current of the solar panel, which is the maximum current the panel can produce when the terminals are shorted.

The parameters we chose to add in the dataset are the peak power, the size, the Vocc, the Vmpp, and the Isc. We think that the paramaters we chose are the most important ones to predict the energy production of a solar panel.

4.1.2.2 Data source

Obtaining datasheets for photovoltaic panels is a crucial step, as they typically reside on manufacturers' websites. However, given the vast number of PV panel models and providers in the market, accumulating this data can be a challenging and time-consuming task. Fortunately, we discovered a more convenient solution – BimSolar! This software contains a library of PV panels, which we leveraged to select a set of panels for our work. The following figure showcases the PV panel library available within BimSolar.

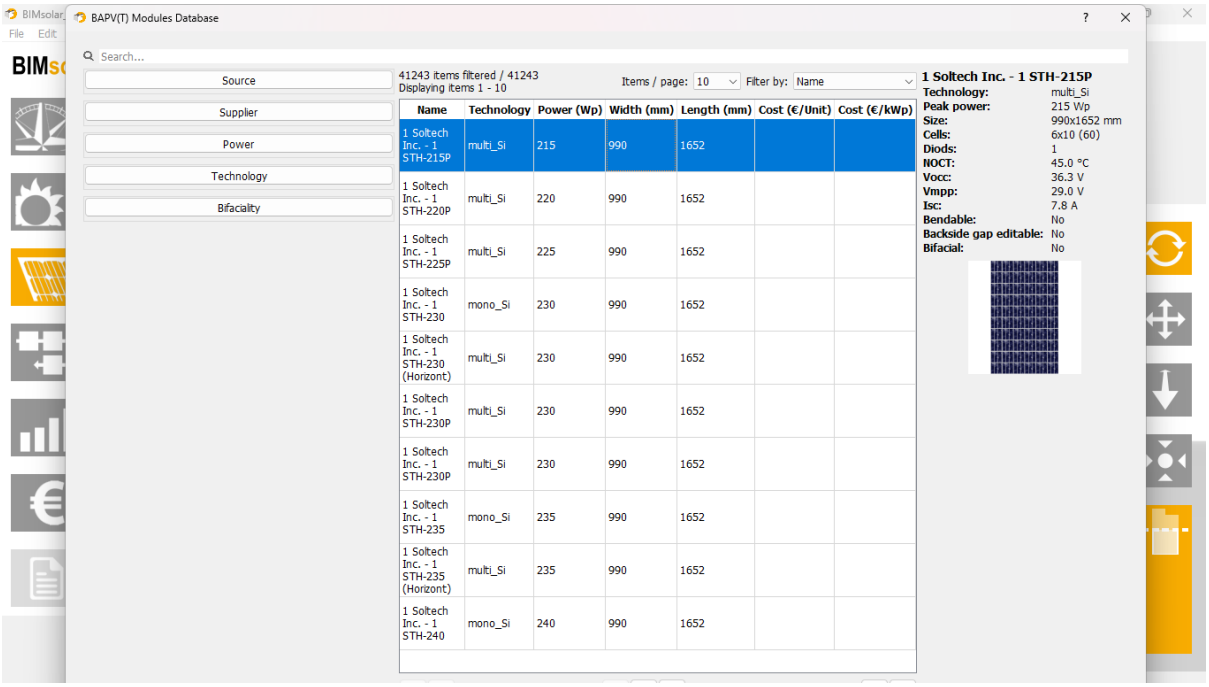

Figure 50: The PV panel library in BimSolar.

The BimSolar library showcased in Figure 50 not only provides access to a vast array of PV panel models but also offers advanced search and filtering capabilities. By leveraging these tools, we can easily navigate through this library, refining our search based on desired characteristics such as supplier, technology type, or power output.

4.1.2.3 Selected solar panels

Thanks to the BimSolar library, we were able to select a diverse set of solar panels for our dataset. This selection process involved choosing panels with varying characteristics.

LR5-54HTH Series: From this series, we selected four solar panels: LR5-54HTH-420M, LR5-54HTH-415M, LR5-54HTH-425M, and LR5-54HTH-430M.

EVPV Series: From this series, we selected four solar panels: EVPV370, EVPV380, EVPV400H, and EVPV410H.

Meyer Series: From this series, we selected five solar panels: Meyer375, Meyer380, Meyer385, Meyer390, and Meyer395.

Blackstar Series: From this series, we selected two solar panels: Blackstar370, and Blackstar420.

Solitek Series: From this series, we selected two solar panels: Solitek400, and Solitek410.

In total, we have selected 17 solar panels with varying characteristics, ensuring a diverse dataset that captures the nuances of different panel types. The following table provides an overview of the selected panels and their key characteristics.

Table 2: The selected solar panels and their key characteristics.

Solar panel	Peak power (Wc)	Size (mm)	V_{occ} (V)	V_{mpp} (V)	I_{sc} (A)
LR5-54HTH-415M	415	1134 x 1722	38.6	32.2	13.9
LR5-54HTH-420M	420	1134 x 1722	38.7	32.4	14.0
LR5-54HTH-425M	425	1134 x 1722	38.9	32.6	14.1
LR5-54HTH-430M	430	1134 x 1722	39.1	32.8	14.1
EVPV370	370	1016 x 1721	44.2	37.4	10.6
EVPV380	380	1016 x 1721	44.3	38.2	10.6
EVPV400H	400	1016 x 1821	48.8	42.1	10.3
EVPV410H	410	1016 x 1821	49.0	42.7	10.4
Meyer375	375	1041 x 1767	44.8	37.8	10.6
Meyer380	380	1041 x 1767	44.5	38.2	10.6
Meyer385	385	1041 x 1767	44.6	38.5	10.6
Meyer390	390	1041 x 1767	44.6	38.9	10.6
Meyer395	395	1041 x 1767	44.8	39.2	10.6
Blackstar370	370	1016 x 1782	40.5	34.9	11.2
Blackstar420	420	1034 x 1722	37.7	32.2	13.5
Solitek400	400	1134 x 1722	37.4	31.3	13.6
Solitek410	410	1134 x 1722	37.4	31.3	13.8

4.1.3 BIM data

Building Information Modeling (BIM) data is a crucial component of our dataset, providing essential information about the buildings. This section outlines the data we need from the buildings, how do we acquire it, and the challenges we faced during the process.

4.1.3.1 Data needed

In the new dataset that we are working on, we need to know the area of the surface where solar panels are going to be placed. This is important because the area of the surface will determine the number of solar panels that can be installed. We also need to know where the solar panels are going to be placed, whether on the roof or the facade. This is important because the energy production of solar panels on the roof is generally higher than those on the facade.

After determining in which surface the solar panels are going to be placed, we need to know how much of that surface will be covered by the solar panels. This is important because the energy production of solar panels is directly proportional to the surface area they cover.

4.1.3.2 Data acquisition

To obtain the data mentioned above, we use IFC files, which are files containing all the necessary information about the buildings. We use the IFC files to extract the area of the surface where the solar panels are going to be placed, this was well explained in the previous chapter.

What if we don't have the IFC files? In this case, we can use drones to take pictures of the buildings and then use photogrammetry software to create 3D models of the buildings or use segmentation algorithms to extract the surface we want to work with. This process is time-consuming and expensive, but it is a viable option when IFC files are not available.

When Industry Foundation Classes (IFC) files are unavailable, drones equipped with cameras or LiDAR sensors can be used to capture building data for photogrammetry, generating point clouds that can be converted into 3D models. Figure 51 showcases two such drones, the DJI Avata 2 and the DJI Mavic 3 Pro, capable of undertaking this task.

Figure 51: Drones that can be used for photogrammetry [15].

Figure 52 provides an example of a point cloud generated by a drone. This detailed representation of the building's structure can then be processed to create a 3D model.

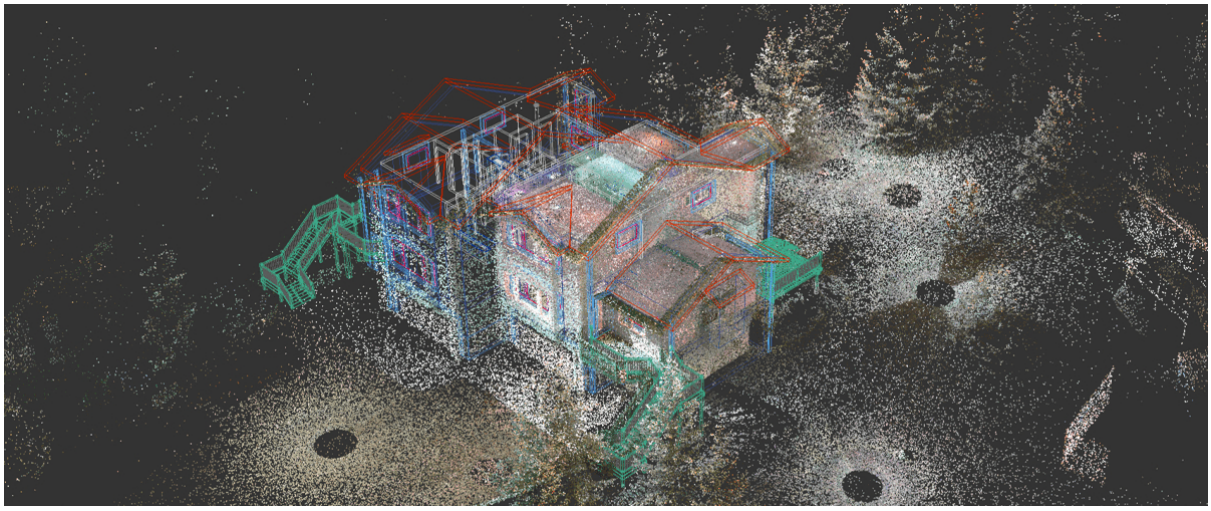

Figure 52: An example of a point cloud generated by a drone.

We were lucky because we had access to IFC files for different buildings in Strasbourg, and Toulouse. We also experimented with using drones to takes pictures of the CESI building to generate the point cloud.

4.1.3.3 Challenges

The biggest challenged we faced while working with IFC files was the lack of standardization in the files. This made the process of extracting the data automatically with a script difficult because in each IFC file, the surfaces will have different identifiers. This is why we had to manually extract the data from the IFC files.

For photogrammetry, the biggest challenge was the processing of the data obtained from the drones. The point clouds generated by the drones are usually very large and require powerful computers to process them. This is why we didn't use this method in our work.

4.2 Data overview

After explaining the process of collecting the data, an overview of the dataset will now be presented. The focus will be on the variability of the data, the number of buildings, the number of features, the number of rows, and other relevant details.

4.2.1 Variability

The dataset we collected is highly diverse, encompassing a wide range of buildings, solar panels, and different percentage of surface area covered by solar panels. This diversity is crucial for training our model, as it ensures that the model can generalize well to unseen data.

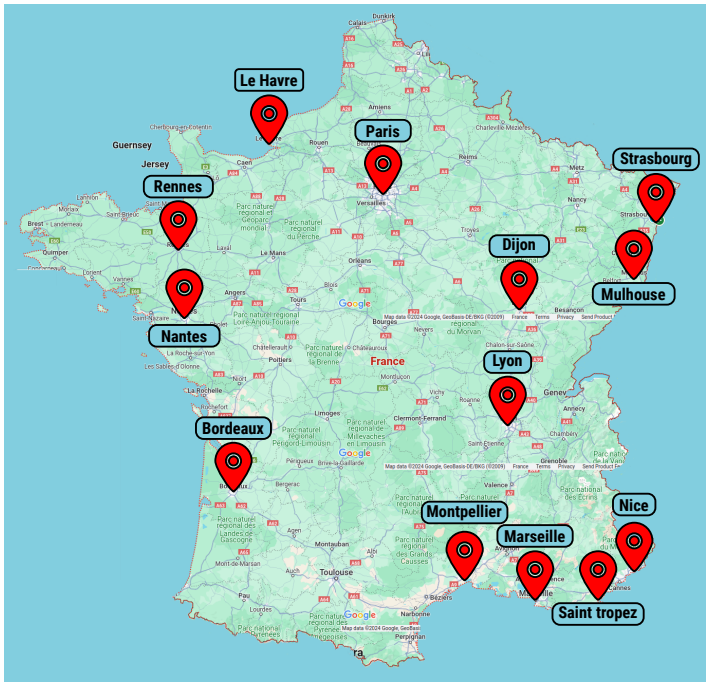

Figure 53: A map of France showing the cities used in the dataset.

Figure 53 shows a map of France with the cities used in the dataset. The dataset includes buildings from Strasbourg, Toulouse, Paris, and other cities in France. We varied the number of cities used in the dataset to capture different distributions of meteorological data, such as temperature, humidity, and solar irradiance. For cities that have generally a higher temperature, the energy production of solar panels will be higher, and vice versa. This diversity in the dataset ensures that the model can generalize well to buildings in different locations.

Alongside the geographical diversity, the dataset also includes a wide range of solar panels with different characteristics, different surfaces, and different exploitation rates. The general pipeline is described in the following figure.

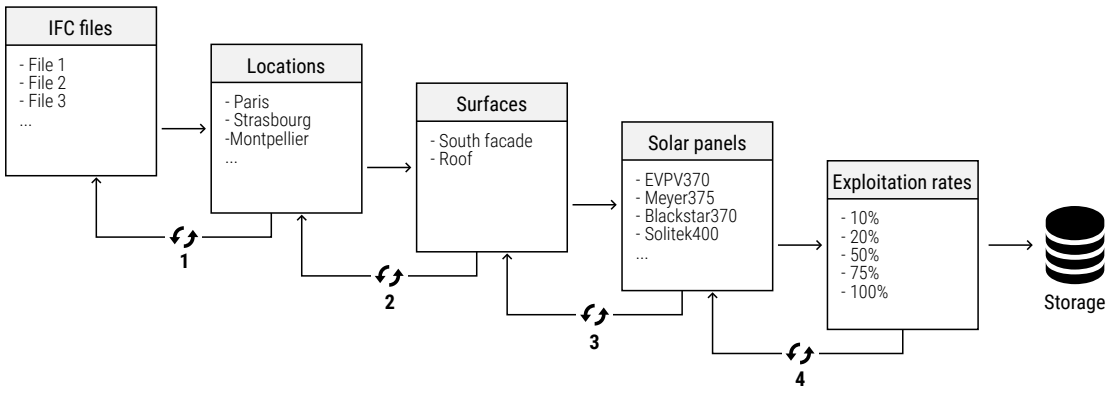

Figure 54: The general pipeline of the synthetic data generation.

In this pipeline, we can see that we start by loading an IFC file, placing the building somewhere in France, so that BimSolar knows where to fetch the meteorological data

from, then we choose a surface of the building where we are going to place the solar panels, after that we choose a type of solar panels to be placed in that selected surface, and finally we choose the percentage of the surface that will be covered by the solar panels.

This process is repeated for every combination of building, location, surface, solar panel, and percentage of surface covered by solar panels. This results in a highly diverse dataset that captures the variability of real-world scenarios. At the end of each process, we obtain the input features as well as the daily energy production.

4.2.2 Data size

Let's use some numbers to see how much data we could have generated just by following the pipeline described in the previous section. We have 14 cities in France, 17 solar panels, 2 surfaces (roof and facade), and 4 percentages of surface covered by solar panels. This results in a total of 14 x 17 x 2 x 4 = 1904 combinations. If we generate data for 365 days, we will have a total of 1904 x 365 = 694,960 rows in the dataset and this is just for one IFC file.

This is great but the process of generating the data is time-consuming and is done manually. This is why we generated just over 100,000 rows in the dataset, which is still a large dataset that captures a huge range of scenarios.

Figure 55 presents an overview of the synthetic dataset. The dataset comprises 101,121 rows and 24 columns. Six columns encompass solar panel characteristics, while 13 columns represent meteorological data. Additionally, two columns contain building-related information, and one column captures energy production data. It is noteworthy that the total number of columns does not sum up to 24, as certain columns will be dropped during the modeling phase.

	Year	Month	Day	Dry Bulb Temperature	Dew Point Temperature	Relative Humidity	Atmospheric Station Pressure	Direct Normal Radiation	Diffuse Horizontal Radiation	Direct Normal Illuminance	...	pv_width	pv_height	pv_vocc	pv_vmpp
0	2017.0	1.0	1.0	8.937500	5.345833	76.333333	99383.339843	0.000000	20.375000	0.000000	...	1.134	1.722	39.1	32.8
1	2017.0	1.0	2.0	5.741667	3.483333	85.666667	99300.000000	0.000000	15.083333	0.000000	...	1.134	1.722	39.1	32.8
2	2017.0	1.0	3.0	7.425000	6.345833	93.166667	99300.000000	0.000000	11.625000	0.000000	...	1.134	1.722	39.1	32.8
3	2017.0	1.0	4.0	4.529167	3.150000	90.791667	99483.337239	0.208333	22.458333	0.125000	...	1.134	1.722	39.1	32.8
4	2017.0	1.0	5.0	2.512500	-0.645833	80.458333	99891.669271	139.416667	20.625000	100.708333	...	1.134	1.722	39.1	32.8
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
101116	2017.0	12.0	27.0	2.737500	0.187500	84.083333	101416.673177	129.875000	29.208333	120.291667	...	1.016	1.721	44.2	37.4
101117	2017.0	12.0	28.0	6.558333	2.000000	73.541667	101387.506835	124.208333	32.916667	88.833333	...	1.016	1.721	44.2	37.4
101118	2017.0	12.0	29.0	10.383334	5.787500	73.916667	101462.502279	162.166667	17.500000	161.875000	...	1.016	1.721	44.2	37.4
101119	2017.0	12.0	30.0	11.575000	7.991667	83.083333	101116.670573	37.916667	31.875000	23.041667	...	1.016	1.721	44.2	37.4
101120	2017.0	12.0	31.0	9.720833	5.766667	76.916667	101162.502929	64.333333	36.416667	46.916667	...	1.016	1.721	44.2	37.4

101121 rows × 24 columns

Figure 55: An overview of the synthetic dataset.

4.3 Modeling

This section focuses on the prediction problem and its potential solution through machine learning models. We'll examine methods for measuring carbon emissions generated during model training. Additionally, we'll explore exposing our trained models to various study case scenarios to validate their generalizability.

4.3.1 The measurement of carbon emissions

Achieving high performance in machine learning models is important, but developing energy-efficient, lightweight models with minimal energy consumption is even more critical. In our work, we prioritize this aspect due to current environmental challenges and our commitment to reducing our impact on the planet. To support this effort, we will utilize a third-party library called CodeCarbon [24] to gather information about energy usage and carbon dioxide emissions associated with our models.

The importance of tracking the environmental impact of training neural networks has been highlighted in recent research. For example, a study by Sevilla Martínez et al. [25] on the CO_2 impact of convolutional network model training for autonomous driving found that "it is necessary to prioritize more energy-efficient algorithms, and that this efficiency should be taken into account as a metric of model quality." The researchers used CodeCarbon to measure the carbon emissions associated with their training process, demonstrating the utility of this tool for quantifying the environmental footprint of machine learning models.

Similarly, a paper by Vera Sousa et al. [26] on the robustness of ConformalLayers to image corruptions also leveraged CodeCarbon to assess the CO_2 emission rates of their deep learning models. The researchers found that the ConformalLayers-based networks had lower emission rates compared to their conventional counterparts, particularly as the network depth increased. This highlights the potential for architectural choices to impact the environmental sustainability of software systems.

Lucía Bouza et al. [26] conducted a comprehensive evaluation of various tools and methods for estimating the energy consumption and carbon footprint of deep learning model training. Their findings shed light on the accuracy and trade-offs of different approaches. Specifically, they reported that the CodeCarbon tool demonstrated high accuracy in estimating energy consumption, with its estimates closely matching the measurements obtained from a wattmeter. However, the authors also noted a trade-off in terms of the additional energy consumption introduced by the monitoring tools themselves, stating that the use of multiple trackers resulted in approximately 7.4% higher energy consumption compared to running the experiment without any trackers. Not only that but the code was almost 10% slower when using trackers than the one without any trackers.

To better understand the emissions generated by our project, we integrated the CodeCarbon library into our codebase. CodeCarbon is a lightweight Python library that tracks the energy consumption and carbon emissions of a Python program. It does this by periodically measuring the power usage of the underlying hardware, including CPUs, GPUs, and RAM. CodeCarbon then calculates the total carbon emissions based on the carbon intensity of the local electricity grid, which it obtains from public data sources on the energy mix of different regions and cloud providers.

Figure 56 illustrates the straightforward methodology used by CodeCarbon. By multiplying the energy consumed (in kWh) by the carbon intensity of the electricity (in gCO_2/kWh), the library is able to provide accurate estimates of the total CO_2 emissions. In cases where specific regional data is not available, CodeCarbon falls back on a global average carbon intensity value.

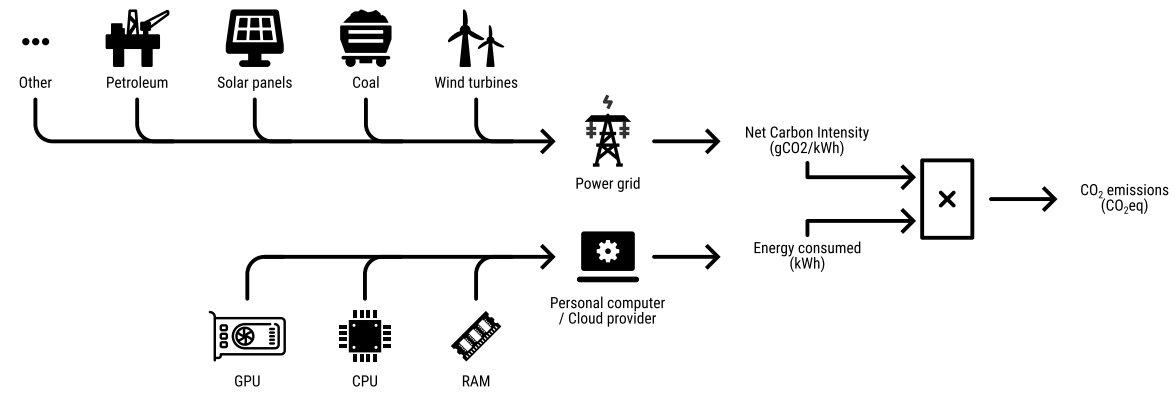

Figure 56: The inner workings of the CodeCarbon library.

While CodeCarbon provides a valuable tool for tracking emissions, it's important to acknowledge its limitations and recognize that it is not the only solution available to developers. Like any tool, CodeCarbon has its own set of assumptions and approximations. For instance, when tracking the energy usage of a CPU that is not present in its database, CodeCarbon relies on a constant value for its calculations. Similarly, if precise information about the carbon intensity of the local electricity grid is unavailable, CodeCarbon falls back on global average values. Despite these caveats, CodeCarbon remains a powerful and accessible option for quantifying the environmental impact of software development as stated by Lucía Bouza et al. [27].

4.3.2 Structuring the data

Contrary to the forecasting problem, we can structure the data differently for the prediction problem. In the forecasting problem, we used the past energy production to predict the future energy production. In the prediction problem, we can only use what we have at the moment to predict the energy production of the day. This means that will split the dataset into two parts, the input features (X) and the target variable (Y).

The dataset is shuffled before splitting it into two parts train, and test. We do this to make sure that we maintain the diversity of the dataset in each part. The dataset is then normalized before feeding it to the models. The training split contains 95% of the dataset, which corresponds to 96,034 rows, while the test split contains 5% of the dataset, which corresponds to 5,055 rows. We have 20 input features used by the models to predict the energy production of the solar panels.

4.3.3 Machine learning models

Our experimentation in predicting solar panel energy production employed various machine learning models, including Random Forest, Gradient Boosting, and XGBoost. A concise overview of each model and its functionality follows.

4.3.3.1 Random Forest

The Random Forest algorithm, adaptable for both classification and regression tasks, is employed here in its regression form. This choice aligns with our goal of predicting a continuous numerical value.

Random Forest is a powerful ensemble learning algorithm that works by constructing multiple decision trees on random subsets of the training data. Each decision tree is built using a different random sample of the data and a random subset of the features. This process of random sampling and feature selection introduces randomness into the model, which helps to reduce overfitting and improve generalization.

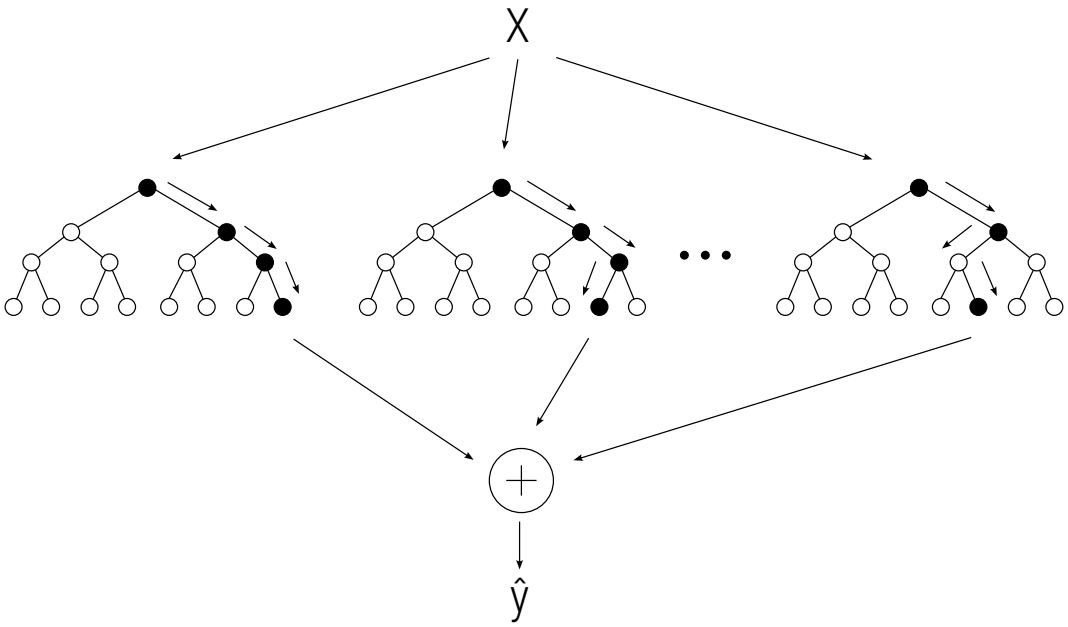

Figure 57: The architecture of the Random Forest algorithm.

During the prediction phase, the Random Forest Regressor takes the input data and passes it through each of the decision trees in the ensemble, see figure 57. Each tree produces a prediction, and the final prediction is obtained by taking the average of all the individual tree predictions. This averaging process helps to reduce the variance and bias of the model, leading to more accurate and robust predictions.

4.3.3.2 Gradient Boosting

Gradient Boosting Regressor is a powerful machine learning technique that combines the principles of gradient descent optimization and boosting. It is an ensemble learning method that builds a strong predictive model by combining multiple weak models, typically decision trees.

The core idea behind Gradient Boosting Regressor is to iteratively train a sequence of decision trees, where each new tree is trained to predict the residual errors made by the previous trees. This process is known as boosting, and it allows the model to gradually improve its predictions by focusing on the instances that were previously difficult to predict. This process is clearly illustrated in figure 58.

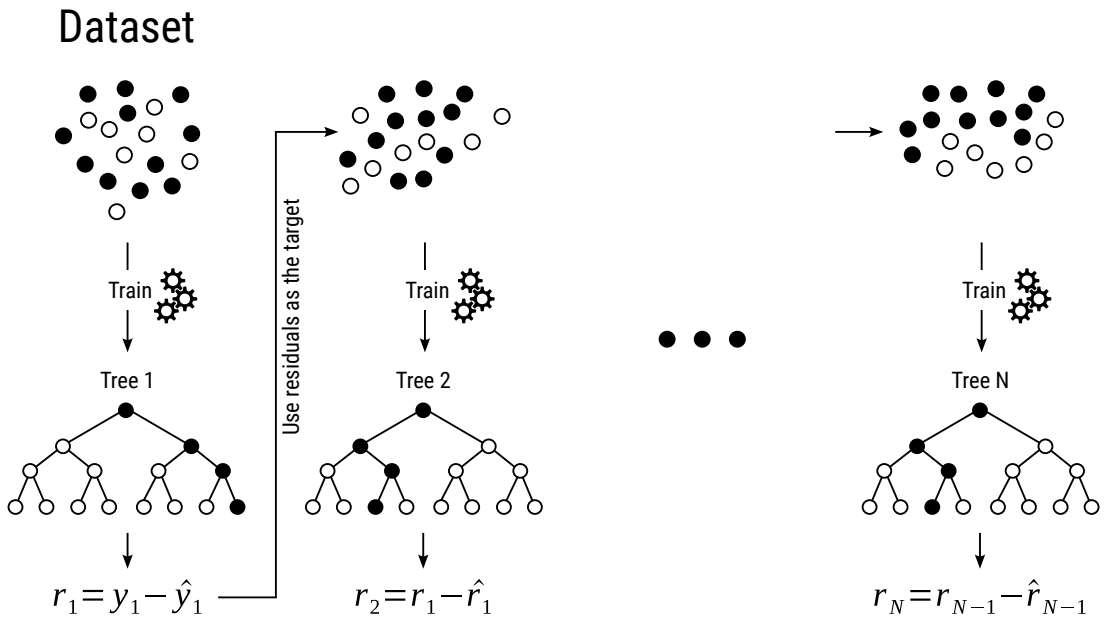

Figure 58: The architecture of the Gradient Boosting algorithm.

Let's explain how the Gradient Boosting Regressor works. The algorithm starts by fitting a simple model to the data, such as a decision tree, this model serves as the initial prediction. It then calculates the residuals (r), which are the differences between the predicted values and the actual values ($y - \hat{y}$). The next step is to fit a new model to the residuals, and this process is repeated for a specified number of iterations or until a stopping criterion is met. The final prediction is obtained by summing the predictions of all the individual models.

4.3.3.3 XGBoost

XGBoost (Extreme Gradient Boosting) is an optimized distributed gradient boosting library designed to be highly efficient, flexible, and portable. It is an implementation of the Gradient Boosting algorithm that has been optimized for speed and performance. The XGBoost algorithm was developed by Tianqi Chen and Carlos Guestrin in 2016 [28] and has become one of the most popular and effective algorithms for structured or tabular data. The XGBoost algorithm works for both Classification and regression tasks, and has the following features and advantages:

- **Regularization:** XGBoost has built-in L1 (Lasso) and L2 (Ridge) regularization to prevent overfitting.
- **Parallelization:** XGBoost is highly optimized for parallel processing and can take advantage of multiple CPU cores to speed up training.
- **Tree Pruning:** XGBoost uses a technique called tree pruning to remove splits that do not provide any significant improvement in the model's performance.
- **Cross-validation:** XGBoost supports cross-validation to evaluate the model's performance and tune hyperparameters.

- **Missing Values:** XGBoost can automatically handle missing values in the data, which simplifies the data preprocessing step.

4.4 Training & results

4.4.1 Training

We trained five machine learning models and one deep learning model on the synthetic dataset. The models we used were Random Forest, Gradient Boosting, XGBoost, Linear Regression, and a dense neural network. The models were trained on the training split of the dataset which contains 96,034 rows and evaluated on the test split which contains 5,055 rows.

To assess the performance of the models, we used four metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), R^2, and the Mean Absolute Percentage Error (MAPE). Alongside these metrics, we also measured the carbon emissions generated by training the models using the CodeCarbon library. These metrics will help us choose a model that is both accurate and energy-efficient.

4.4.2 Results

The performance of the models was evaluated on the test split of the dataset using various metrics, including Mean Squared Error (MSE), R-squared (R^2), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). As presented in Table 3, the Random Forest and XGBoost models outperformed the other models across all metrics, demonstrating high accuracy and efficiency in predicting solar panel energy production. The XGBoost model achieved the best overall performance, with an MSE of 2316.66, an R^2 score of 0.987233, an MAE of 16.01, and an MAPE of 0.181846.

Table 3: The performance of the models on the test set.

Metrics	**RF**	**LR**	**GBR**	**XGBR**	**DNN**
MSE	3340.03	51509.02	4942.67	**2316.66**	5725.44
R^2	0.981593	0.716131	0.972761	**0.987233**	0.968447
MAE	**12.99**	155.33	30.42	16.01	33.37
MAPE	**0.032495**	10.671123	0.498437	0.181846	1.120349

In addition to predictive performance, the carbon emissions and energy consumption during model training were assessed, as shown in Table 4. The XGBoost model proved to be the most energy-efficient, taking only 0.11 seconds to train and consuming negligible amounts of energy and generating minimal carbon emissions ($9.06 * 10^{-6}$ gCO_{2eq}). In contrast, the Dense neural network (DNN) model was the least efficient, requiring 202.98 seconds of training time and consuming 0.861 Wh of energy, resulting in higher carbon emissions of 0.017 gCO_{2eq}.

Table 4: The carbon emissions generated by training the models.

Metrics	**RF**	**LR**	**GBR**	**XGBR**	**DNN**
Duration (s)	103.45	31.28	0.854	**0.110**	202.98
Emissions (gCO_{2eq})	0.009	0.003	7.24e-5	**9.06e-6**	0.017
Energy consumed (Wh)	0.439	0.133	3.62e-3	**4.53e-4**	0.861

The numbers indicate that the models are doing good but let's see a plot of the predictions of the XGBoost model compared to the actual values. The following figure shows the actual values of the energy production of the solar panels and the predictions of the XGBoost model.

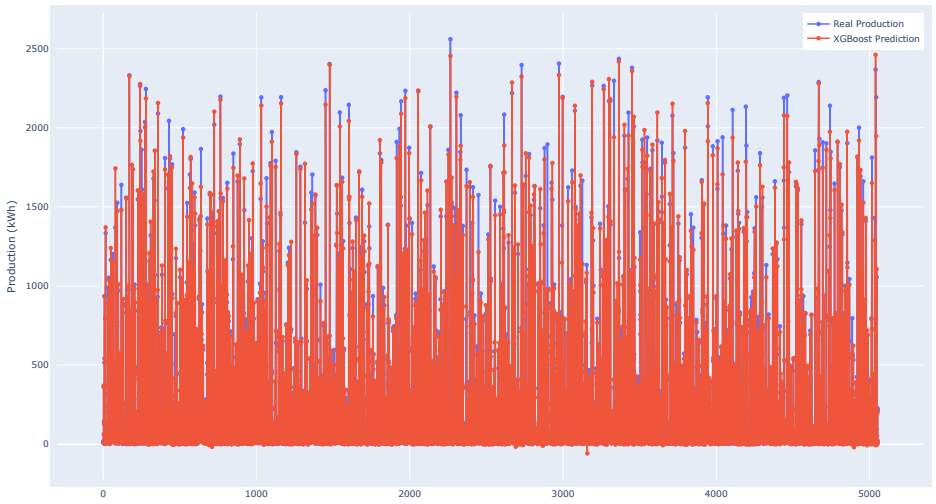

Figure 59: The actual values and the predictions of the XGBoost model.

The plot in figure 59 shows that the XGBoost model is doing a good job at predicting the energy production of the solar panels. The predictions are very close to the actual values, but sometimes the model gives negative values for the energy production, which is not possible. This problem occurs because in some days we have negligible production and the model tries to predict a value close to zero, but sometimes it gives a negative value. This problem can be fixed later by processing the predictions and setting the negative values to zero.

4.5 Study cases

To validate the generalization of the models, we exposed them to different study cases scenarios. We have three study cases scenarios, the first one is the CESI building in Strasbourg, the second one is the CCCA-BTP building in Toulouse, and the third one is the Les compagnons du devoir building in Strasbourg. For some buildings we have the real energy production data, and for others we don't, so we used BimSolar to estimate the energy production for those buildings.

Our models will be used to predict the energy production of solar panels on buildings within the study cases. Predictions will be compared to actual energy production values.

This comparison serves to determine how effectively the models generalize to new and unseen data.

4.5.1 Study case 1 - CESI building

We talked about the CESI building before, figure 39 shows the CESI building. We will start by simulating the energy production in BimSolar and then use our models to compare their predictions to BimSolar's predictions.

The following figure shows the CESI building inside BimSolar. The solar panels are located in the south facade of the building. In total we have 48 solar panels, each with a power of 415Wc.

Figure 60: The CESI building inside BimSolar.

The following table shows the performance metrics of the models on the CESI building.

Table 5: The performance of the models on the CESI building.

Metrics	**DT**	**RF**	**GBR**	**LR**	**XGBR**
MSE	120.75	122.09	191.89	18252.70	**52.34**
R^2	0.787914	0.785557	0.662966	-31.058978	**0.908071**
MAE	8.61	8.69	10.27	117.47	**5.13**
MAPE	0.23	0.24	0.46	12.59	**0.18**

From the table, we can see that the XGBoost model outperformed the other models on the CESI building, achieving the lowest MSE, the highest R^2 score, the lowest MAE, and the lowest MAPE.

The following figure shows the predictions of the XGBoost model compared to BimSolar's predictions on the CESI building.

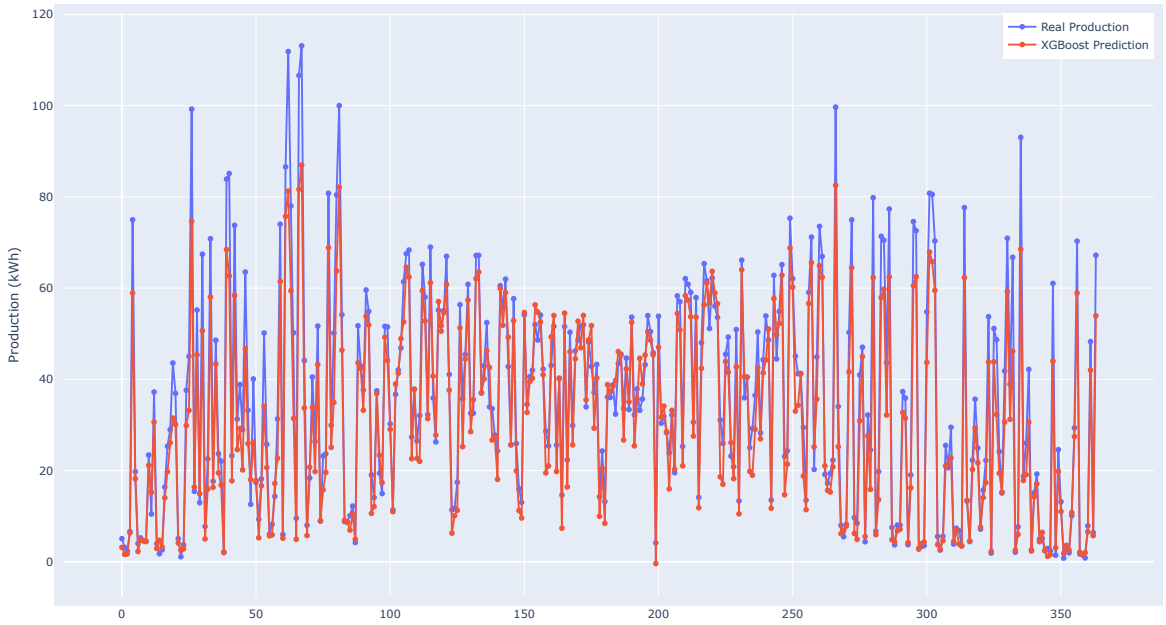

Figure 61: The predictions of the XGBoost model compared to BimSolar's predictions on the CESI building.

4.5.2 Study case 2 - Les compagnons du devoir

Les compagnons du devoir is a building located in Strasbourg. This building started to produce energy from solar panels in the beginning of March 2024. The following figure shows the building and the solar panels on the roof.

Figure 62: Les compagnons du devoir building in Strasbourg as viewed from Google Maps.

We created a 3D model of the building and imported the file in BimSolar after that we placed the solar panels on the roof of the building. The following figure shows the building and the solar panels in BimSolar. In total, we have 156 solar panels, each with

a power of 415Wc.

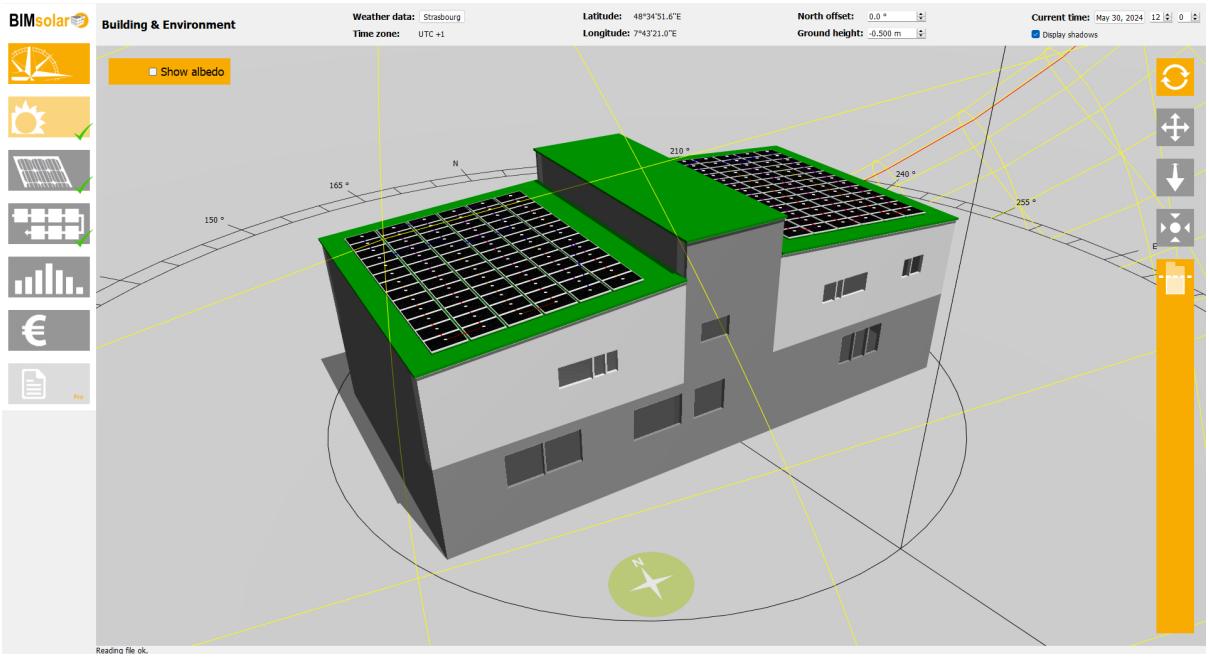

Figure 63: Les compagnons du devoir building in BimSolar.

Mirroring the approach taken with the CESI building, our models will be applied to predict energy production for solar panels on the Les compagnons du devoir building. Model predictions will subsequently be compared against actual energy production values. The table below presents the performance metrics of the models for the Les compagnons du devoir building.

Table 6: The performance of the models on the Les compagnons du devoir building.

Metrics	**DT**	**RF**	**GBR**	**LR**	**XGBR**
MSE	**14.23**	55.30	7521.01	117562.30	856.38
R^2	**0.999934**	0.999742	0.964975	0.452519	0.996012
MAE	**0.52**	5.78	68.03	282.67	22.57
MAPE	**0.000666**	0.008107	0.134309	0.716552	0.043362

From the table, we can see that the Decision Tree model outperformed the other models on the Les compagnons du devoir building, achieving the lowest MSE, the highest R^2 score, the lowest MAE, and the lowest MAPE.

The following figure shows the predictions of the Decision Tree model compared to BimSolar's predictions on the Les compagnons du devoir building.

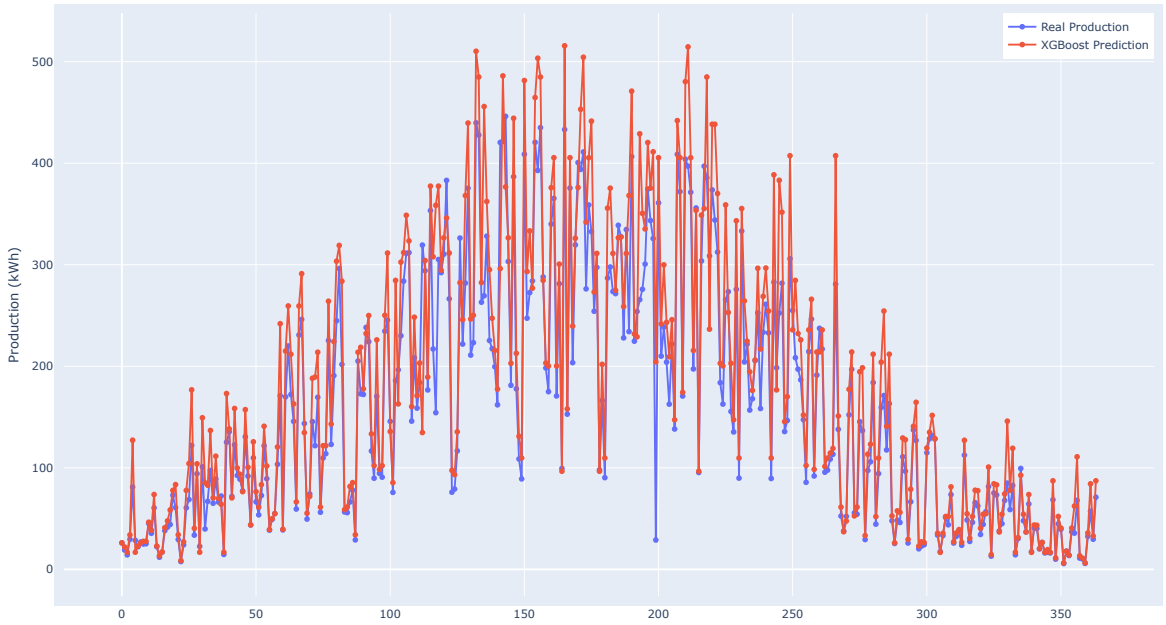

Figure 64: The predictions of the Decision Tree model compared to BimSolar's predictions on the Les compagnons du devoir building.

4.5.3 Study case 3 - CCCA-BTP Muret

CCCA-BTP is a building located in Toulouse. This building contains two roofs where we placed the solar panels. The following figure shows the building and the solar panels on the roofs. In total, we have roughly 600 solar panels. This number is very high compared to the other buildings, but because we trained the models on a very diverse dataset, we expect the models to generalize well to this building.

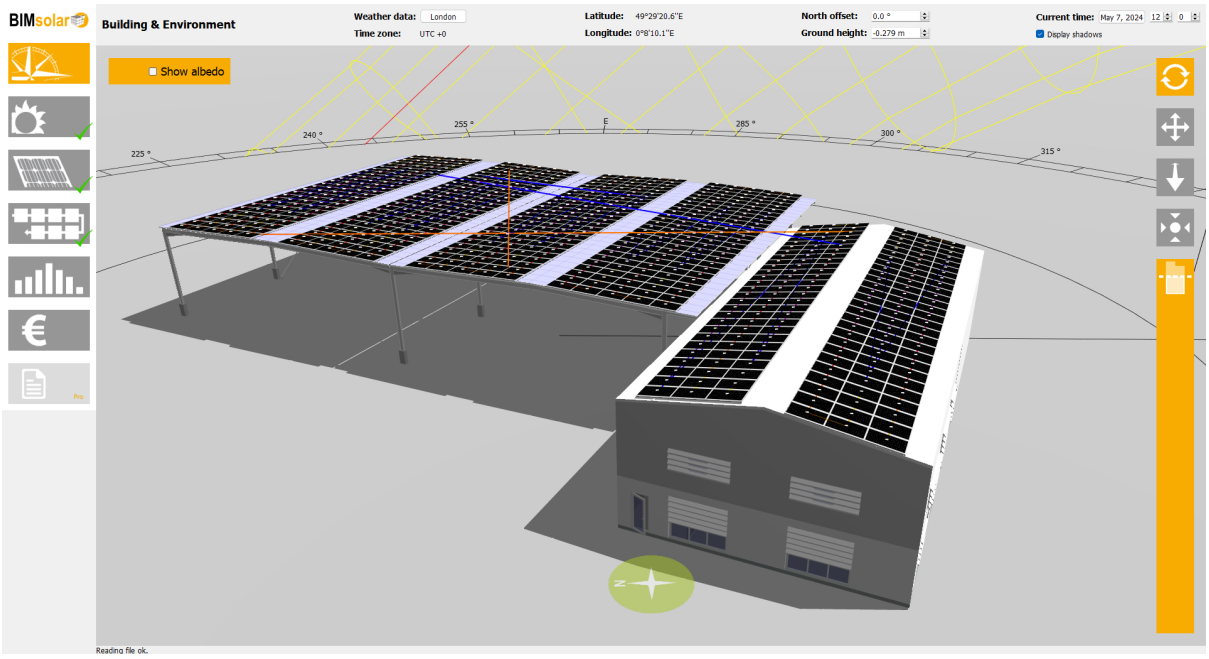

Figure 65: CCCA-BTP building in Mulhouse.

Our models will be applied to predict the energy production of the solar panels on the CCCA-BTP building. Performance will be assessed by comparing these predictions to the actual energy production values. The following table details the performance metrics of the models on the CCCA-BTP building.

Table 7: The performance of the models on the CCCA-BTP building.

Metrics	**DT**	**RF**	**GBR**	**LR**	**XGBR**
MSE	**18.16**	55.04	7294.08	117562.30	856.38
R^2	**0.999915**	0.999744	0.966032	0.452519	0.996012
MAE	**0.52**	5.68	67.26	282.67	22.57
MAPE	**0.000652**	0.008010	0.129535	0.716552	0.043362

From the table, we can see that the Decision Tree model outperformed the other models on the CCCA-BTP building, achieving the lowest MSE, the highest R^2 score, the lowest MAE, and the lowest MAPE.

The following figure shows the predictions of the XGBoost model compared to BimSolar's predictions on the CCCA-BTP building.

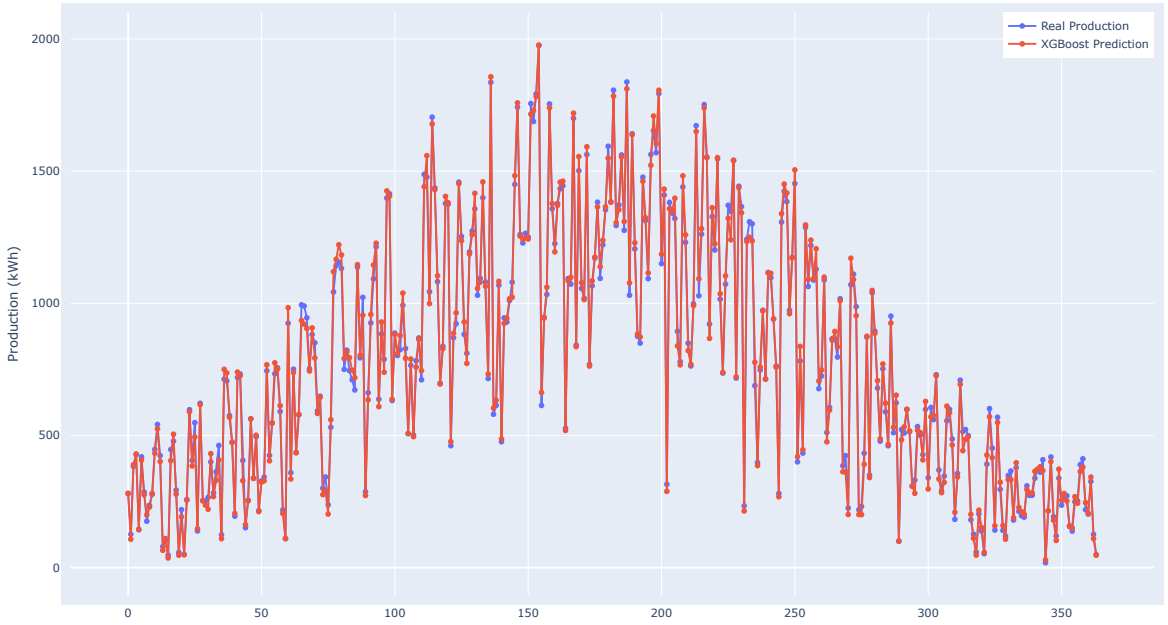

Figure 66: The predictions of the XGBoost model compared to BimSolar's predictions on the CCCA-BTP building.

4.6 Conclusion

In this chapter, we presented the synthetic dataset that we generated using BimSolar. We described the general pipeline of the synthetic data generation and explained how we

structured the data for the prediction problem. We then introduced the machine learning models that we used to predict the energy production of solar panels and discussed the importance of measuring carbon emissions during model training. We trained the models on the synthetic dataset and evaluated their performance using various metrics. The XGBoost and Decision Tree models emerged as the top performers, achieving high accuracy and energy efficiency. We then validated the generalization of the models by exposing them to different study cases scenarios. The models demonstrated strong predictive capabilities across a range of scenarios, highlighting their potential for real-world applications.

Chapter 5

The web application

This chapter showcases our user-friendly and simplistic web application, divided into three sections: the data hub, prediction, and decarbonization. The data hub allows users to generate input data for the model. The prediction section enables forecasting solar panel energy production on buildings. Meanwhile, the decarbonization section provides insights into building energy performance, emission levels, and decarbonization rates. Designed with an intuitive and modern interface, the web application seamlessly guides users through each process step, ensuring a smooth and straightforward experience.

5.1 The data hub

The data hub section allows users to specify a building's geographical coordinates (latitude and longitude) along with desired start and end dates. These inputs are leveraged to generate meteorological data specific to the building's location. Upon clicking the "Generate Data" button, the backend system fetches the required meteorological data from an API. The following figure demonstrates the initial part of the data hub section.

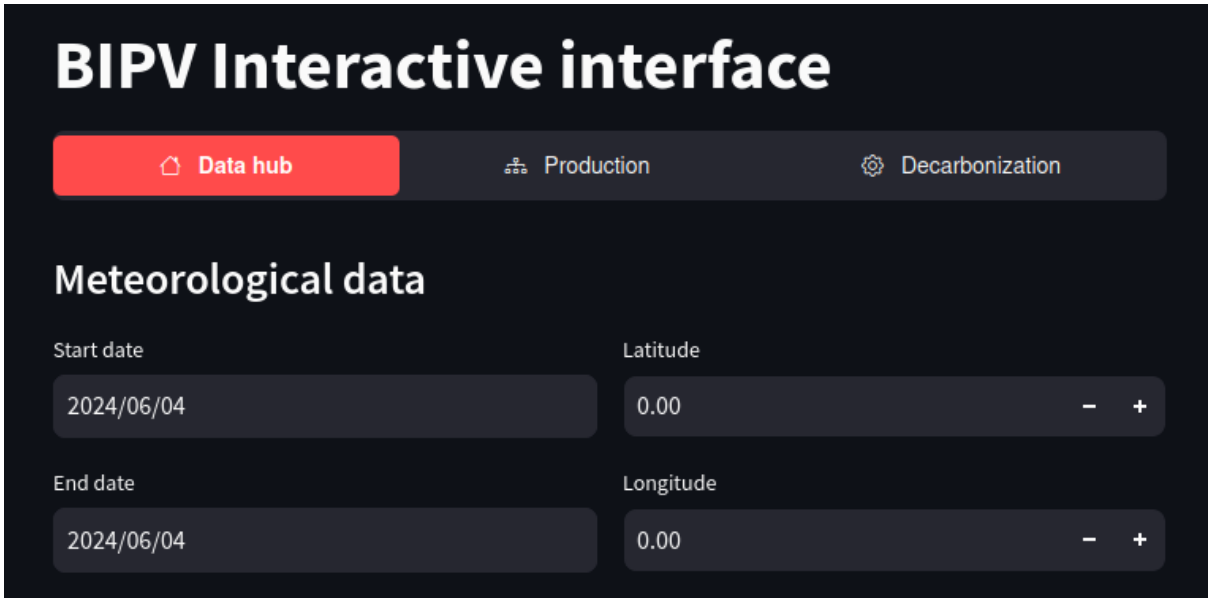

Figure 67: The first part of the data hub section.

Upon completing the meteorological data inputs, users can proceed to select a solar panel. The web application offers a list of existing solar panel options. However, if a user has a specific solar panel in mind, they can leverage the "Use Manual Mode" checkbox

and provide the necessary information. The subsequent figure illustrates the solar panel selection process within the application.

Figure 68: The solar panel selection process in the data hub section.

When checking the "Use Manual Mode" checkbox, the information that the user should specify is show in the following figure.

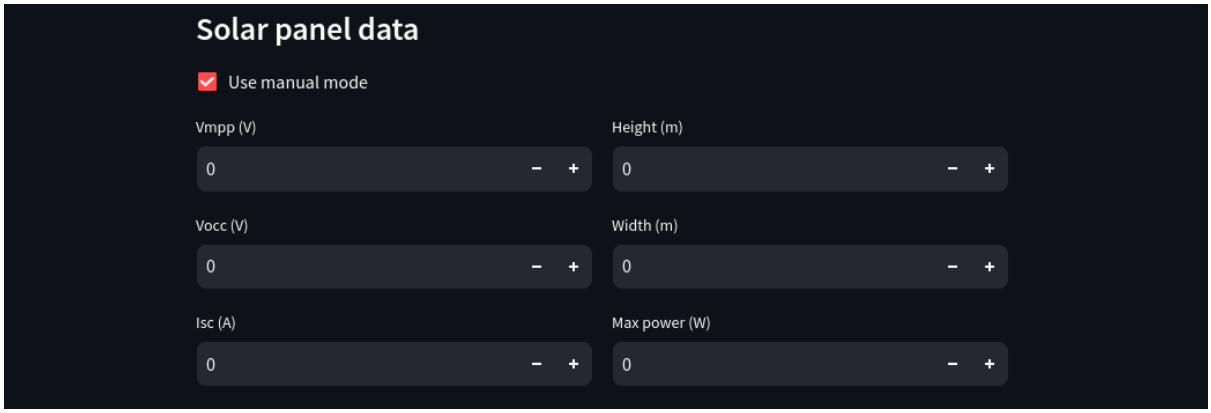

Figure 69: The manual mode for the solar panel selection.

The last part in the data hub section allows the user to select the desired location for solar panel installation on the building. Options include the roof or the south-facing facade. Subsequently, the user specifies the surface area of the chosen facade or roof where the solar panels will be placed. Finally, the user can indicate the exploitation rate, or the fraction of the available surface area to be utilized for solar panel installation. The following figure depicts the final segment of the data hub section, facilitating these selections.

Figure 70: The final part of the data hub section.

Once all the necessary inputs have been provided, users can trigger the data generation process by clicking the "Generate Data" button. The backend system then processes the inputs and generates the necessary meteorological and solar panel data. This data is presented to the user in a tabular format, allowing them to conveniently view and download the data if desired. The following figure illustrates the resulting data table display.

	date	Dry Bulb Temperature	Dew Point Temperature	Diffuse Horizontal Radiation	Direc
0	2024-02-01 00:00:00	5.8085	4.3	49.383	
1	2024-02-02 00:00:00	6.9792	2.7292	21.9792	
2	2024-02-03 00:00:00	7.0208	0.8979	17.4375	
3	2024-02-04 00:00:00	7.3958	0.5917	20.125	
4	2024-02-05 00:00:00	7.0417	0.425	19.3958	
5	2024-02-06 00:00:00	3.1667	0.9917	61.2708	
6	2024-02-07 00:00:00	5.3958	2.4396	31.2708	
7	2024-02-08 00:00:00	8.1458	5.5521	51.6042	
8	2024-02-09 00:00:00	7.375	6.1708	42.3125	
9	2024-02-10 00:00:00	4.7292	4.0083	25.3958	

Figure 71: The data preview in the data hub section.

Now the user can go to the prediction section to predict the energy produced by the solar panels in the range of dates that he specified in the data hub section.

5.2 The prediction section

This section enables users to use a trained model for generating solar panel energy production estimates for the specified building. Additionally, a dropdown menu provides users with the flexibility to change the solar panel selection if desired. The following figure showcases the model and the solar panel selection options.

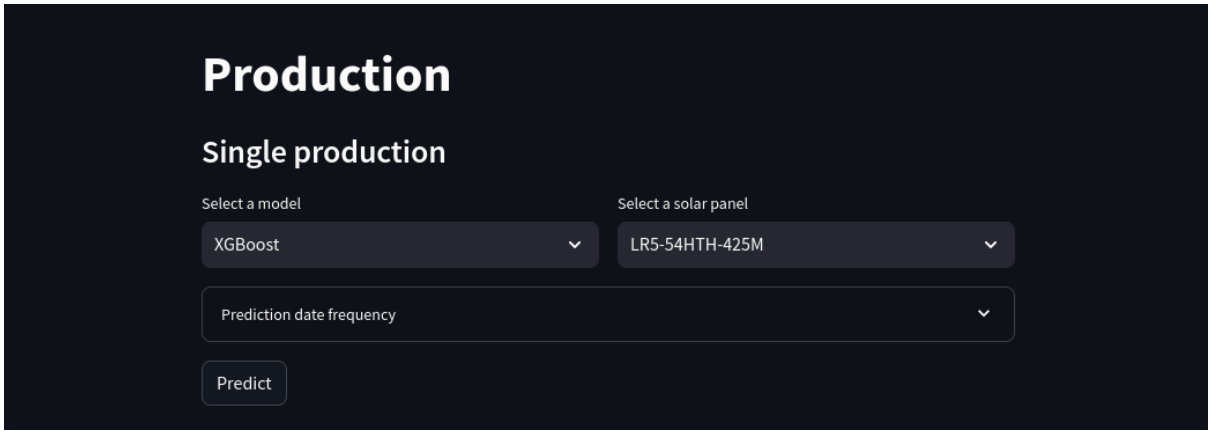

Figure 72: The model and solar panel selection options in the prediction section.

Users can predict solar panel energy output by selecting daily or monthly options under "Prediction date frequency" and clicking the "Predict" button. The system then displays the prediction of energy production for the chosen timeframe, as illustrated in the following figure.

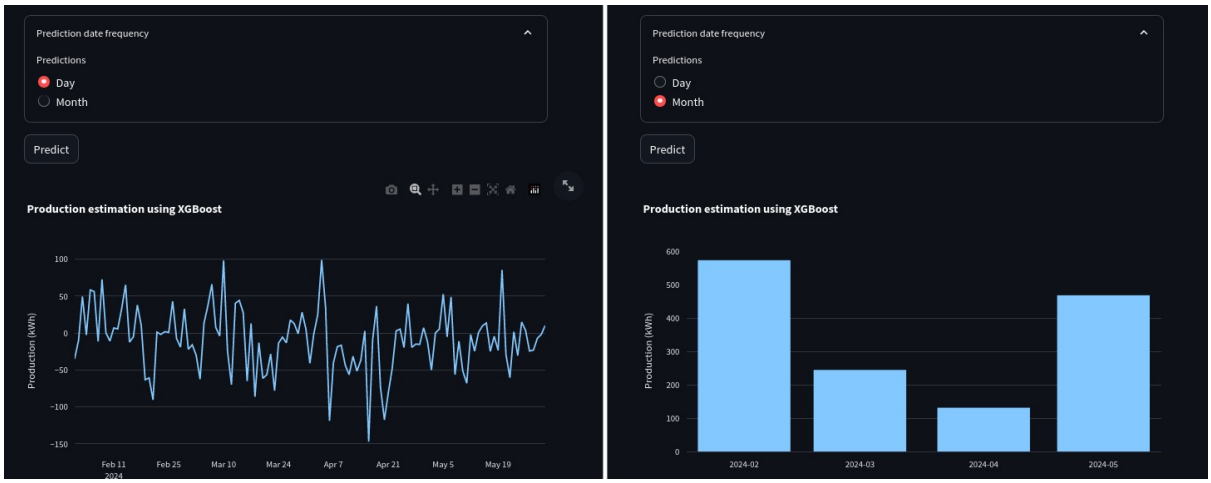

Figure 73: The energy production prediction in the prediction section.

In the same page, users can compare the energy produced between different solar panels. Users can select any number of solar panels in the select box and then click on the "Predict" button. The following figure shows the comparison between three solar panels.

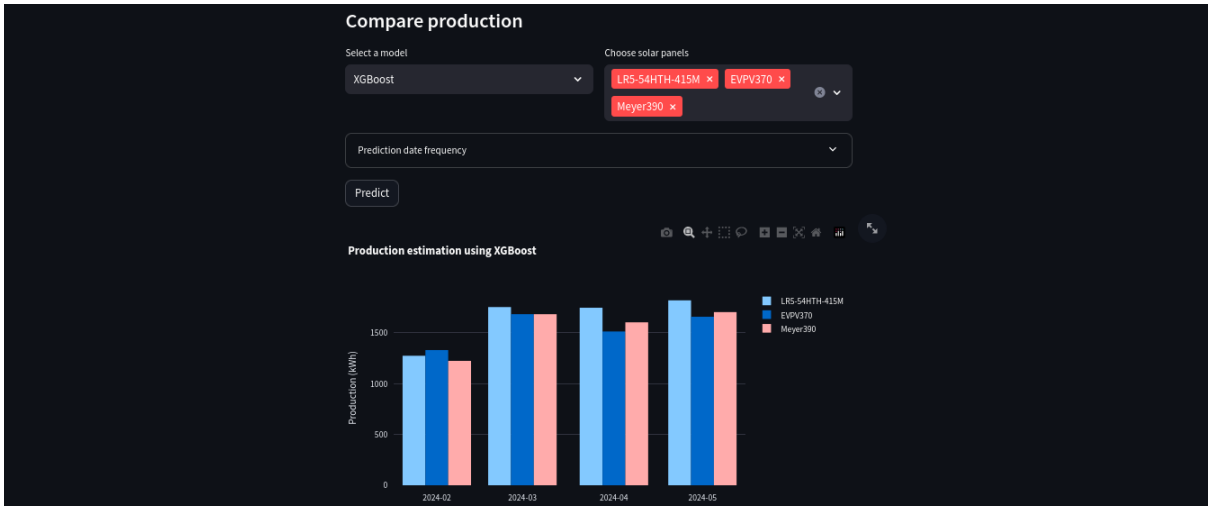

Figure 74: The energy production comparison in the prediction section.

After the user has predicted the energy production of the solar panels, he can go to the decarbonization section to see the energy performance of the building and the carbon emissions generated by the energy consumption of the building.

5.3 The decarbonization section

The decarbonization section starts by asking users to enter their building's annual energy consumption or the consumption for a specific timeframe. They then need to provide the usable surface area of the building. Based on these inputs, the section calculates and displays the building's energy performance, its emission levels, and crucially, its decarbonization rate. The following figure illustrates the initial part of the decarbonization section.

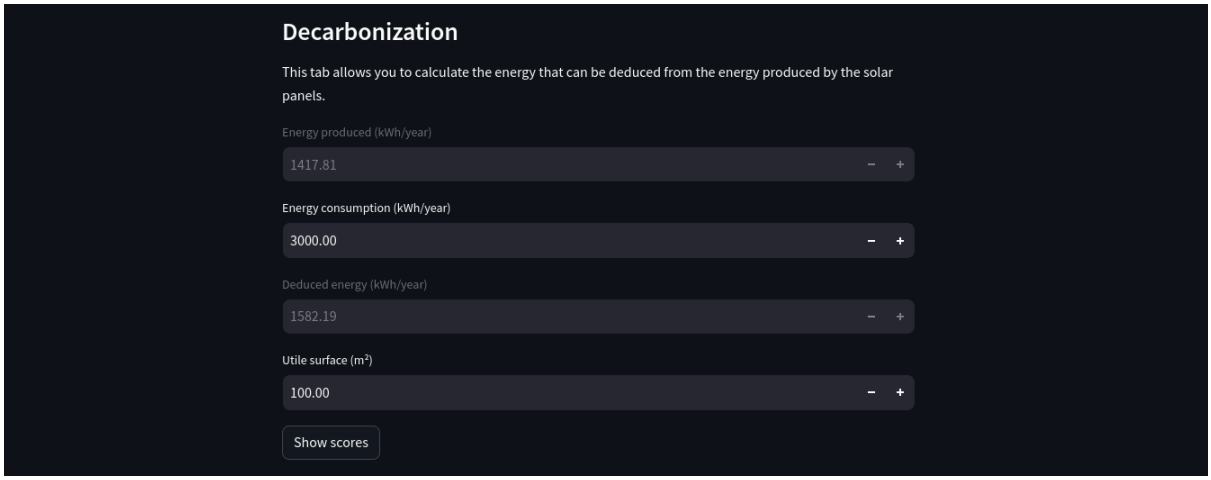

Figure 75: The initial part of the decarbonization section.

Clicking the "Show scores" button reveals the building's energy performance, emissions, and decarbonization rate. The first graph (illustrated in the figure below) assesses overall energy efficiency, ranging from A (most efficient) to G (least efficient). Arrows in the graph indicate how installing solar panels (PV) would impact the building's energy rating.

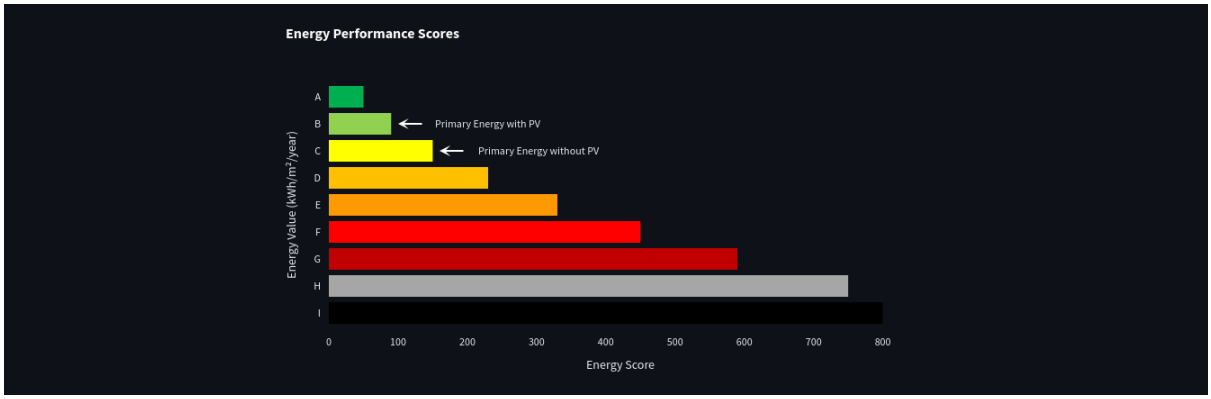

Figure 76: The energy performance score in the decarbonization section.

In this example, we can see that after installing the solar panels the building's energy performance score improved from C to B.

The second graph (shown in the figure below) displays the building's carbon emissions, with the potential reduction in emissions following solar panel installation indicated by arrows.

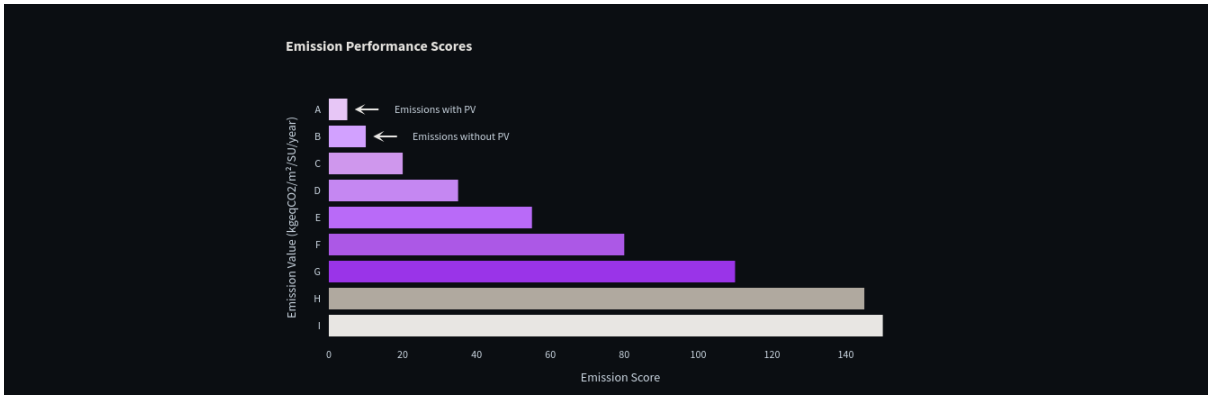

Figure 77: The carbon emissions in the decarbonization section.

In this example, we can see that after installing the solar panels the building's emission performance scores improved from B to A.

Finally, the last information that the user can see is the decarbonization rate of the building. The decarbonization rate is the percentage of carbon emissions that have been reduced by installing the solar panels. The following figure shows the decarbonization rate of the building.

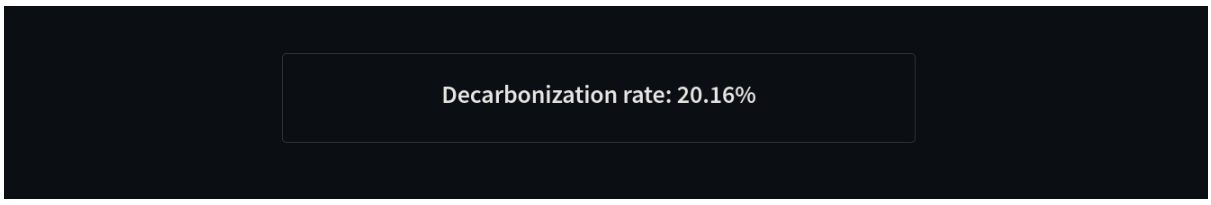

Figure 78: The decarbonization rate in the decarbonization section.

In this example, we can see that the decarbonization rate of the building is 20.16%. To

increase the value of the decarbonization rate, the user can install more solar panels or improve the energy efficiency of the building.

5.4 Conclusion

In this chapter, we presented our web application, which provides users with a comprehensive and user-friendly interface for generating data, predicting solar panel energy production, and assessing building energy performance and decarbonization rates. The application is designed to guide users through each step of the process, from data input to result visualization, ensuring a seamless and intuitive experience. By leveraging the application's functionalities, users can gain valuable insights into their building's energy performance, emission levels, and decarbonization potential, empowering them to make informed decisions and drive sustainable energy practices.

Chapter 6

Summary and future prospects

6.1 Summary and conclusions from current work

This internship project aimed to develop an AI tool capable of accurately predicting the energy production of photovoltaic (PV) systems integrated into buildings. Building upon the research of Youssef & Ilyas, we reimplemented their approach and addressed its limitations by generating a significantly larger and more diverse dataset. We achieved this by leveraging BimSolar to generate synthetic data, incorporating a wider range of building types, solar panel models, and installation locations. This resulted in a dataset ten times larger than the original, ensuring greater accuracy and generalizability of the models. We also implemented an automated data collection process using APIs like NASA POWER and Solcast, simplifying data acquisition and significantly accelerating dataset assembly.

Furthermore, we transitioned from a forecasting problem to a prediction problem, allowing for direct energy production estimation without historical data. This shift offered significant advantages, particularly for replacing sizing software with machine learning models. We integrated the CodeCarbon library to measure the carbon emissions generated by training the models, emphasizing the importance of energy-efficient and lightweight models. We then evaluated the performance of various machine learning models, including Random Forest, Gradient Boosting, XGBoost, and a dense neural network, using relevant metrics like MSE, R^2, MAE, and MAPE. XGBoost and Decision Tree emerged as the top performers, achieving high accuracy and energy efficiency.

Finally, we validated the models on three real-world study cases, showcasing the ability to accurately predict energy production for different buildings with diverse solar panel installations.

6.2 Future prospects

Although this internship project made significant progress, several areas for future research and development remain open. Expanding the dataset by integrating real-world data, especially from diverse countries and regions, would significantly enhance the accuracy and ability of our model to generalize to unseen scenarios. Furthermore, automating data collection processes could be improved. This could involve developing more robust pipelines and potentially exploring methods to automatically export data from BimSolar. Furthermore, we could focus on incorporating factors such as shading losses, the age of the solar panel, a wider range of PV technologies, return on investment calculations, cost

evaluations, and adapt the calculation of the decarbonization rate for each country and region. By including these crucial factors, the model's predictive power and practical applicability would be significantly enhanced.

References

[1] International Energy Agency. *Electricity 2024: Analysis and forecast to 2026*, https://iea.blob.core.windows.net/assets/18f3ed24-4b26-4c83-a3d2-8a1be51c8cc8/Electricity2024-Analysisandforecastto2026.pdf

[2] European Council. Council of the European Union *Fit for 55*.

[3] Réseau de Transport d'Électricité. *Futurs énergétiques 2050*. URL https://assets.rte-france.com/prod/public/2021-10/Futurs-Energetiques-2050-principaux-resultats_0.pdf

[4] Visbox, Inc, *CAVE systems*, http://www.visbox.com/products/cave/

[5] I. Abouelaziz, Y. Jouane, *Photogrammetry and deep learning for energy production prediction and building-integrated photovoltaics decarbonization*, CESI LINEACT, 2023.

[6] CAMS. *Copernicus Atmosphere Monitoring Service*. URL https://www.soda-pro.com/web-services/radiation/cams-radiation-service

[7] OpenDataSoft. URL https://data.opendatasoft.com/pages/home/

[8] DeepLearningBrasilia. *Deep Learning: Recurrent Neural Networks*. Medium, 2018. https://medium.com/deeplearningbrasilia/deep-learning-recurrent-neural-networks-f9482a24d010

[9] S. Hochreiter, J. Schmidhuber, 1997. *Long short-term memory*. Neural Computation, 9(8):1735–1780. URL https://www.bioinf.jku.at/publications/older/2604.pdf

[10] Wikipedia. *Long short-term memory*. https://en.wikipedia.org/wiki/Long_short-term_memory

[11] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, 2014. *Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation*. URL https://arxiv.org/pdf/1406.1078v3

[12] Wikipedia. *Gated recurrent unit*. https://en.wikipedia.org/wiki/Gated_recurrent_unit

[13] BimSolar. *Solar Architecture, Integrated PhotoVoltaics.* URL https://www.bim-solar.com/

[14] NASA POWER. *NASA Prediction Of Worldwide Energy Resources*. URL https://power.larc.nasa.gov/

[15] DJI. URL https://www.dji.com/fr/camera-drones?site=brandsite&from=nav

[16] Abhishek Anand, Renuga A P Verayiah, Muhamad Mansor, Tengku Juhana Tengku Hasim, Amritanshu Shukla, Hitesh Panchal, Atul Sharma, L. Natrayan, Abhinav Kumar. *A comprehensive analysis of small-scale building integrated photovoltaic system for residential buildings: Techno-economic benefits and greenhouse gas mitigation potential*.

[17] *Solar PV Issues, Top Solar Energy System Losses*, https://solarsme.com/top-solar-energy-system-losses/

[18] Mohammad H. Alomari, Ola Younis, Sofyan M. A. Hayajneh. *A Predictive Model for Solar Photovoltaic Power using the Levenberg-Marquardt and Bayesian Regularization Algorithms and Real-Time Weather Data.*

[19] Bowoo Kim, Dongjun Suh. *Solar PV Generation Prediction Based on Multisource Data Using ROI and Surrounding Area.*

[20] X.J. Luo, Lukumon O. Oyedele, Anuoluwapo O. Ajayi, Olugbenga O. Akinade. *Comparative study of machine learning-based multi-objective prediction framework for multiple building energy loads.*

[21] Pamela Ramsami, Vishwamitra Oree. *A hybrid method for forecasting the energy output of photovoltaic systems.*

[22] Abhishek Kumar Tripathi, Mangalpady Aruna, P.V. Elumalai, Krishnasamy Karthik, Sher Afghan Khan, Mohammad Asif, Koppula Srinivas Rao. *Advancing solar PV panel power prediction: A comparative machine learning approach in fluctuating environmental conditions.*

[23] Solcast, 2019. *Global solar irradiance data and PV system power output data.* URL https://solcast.com/

[24] V. Schmidt, K. Goyal, A. Joshi, B. Feld, L. Conell, N. Laskaris, D. Blank, J. Wilson, S. Friedler, S. Luccioni, *CodeCarbon: estimate and track carbon emissions from machine learning computing*, 2021, https://github.com/mlco2/codecarbon.

[25] F. Sevilla Martínez, R. Parada, J. Casas-Roma, *CO2 impact on convolutional network model training for autonomous driving through behavioral cloning*, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2022.

[26] E. V. Sousa, C. N. Vasconcelos, L. A. F. Fernandes, *An analysis of ConformalLayers' robustness to corruptions in natural images*, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2022.

[27] L. Bouza, A. Bugeau, L. Lannelongue, *How to estimate carbon footprint when training deep learning models? A guide and review.* Environmental Research Communications, 2023

[28] T. Chen, & C. Guestrin (2016). *XGBoost: A Scalable Tree Boosting System.* ArXiv, abs/1603.02754.