GREEN-LOOP

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under the Grant Agreement No **101057765**

GREEN LOOP

Sustainable manufacture systems towards novel bio-based materials

WP2 – Sustainability and Circularity by design

D2.11 – GREEN-LOOP Machine learning optimisation

Version 2.0

Document information

Contractual Due date: 30.09.2024	**Delivery Date**: 4.10.2024
Author(s): IDENER	
Lead Beneficiary of Deliverable: IDENER	
Dissemination level: Public	
Nature of the Deliverable: OTHER, accompanied by the present report.	
Internal Reviewers: IDENER	

GREEN LOOP KEY FACTS

Project title	Sustainable manufacture systems towards novel bio-based materials
Starting date	09/01/2022
Duration in months	36
Call (part) identifier	TWIN GREEN AND DIGITAL TRANSITION 2021 (HORIZON-CL4-2021-TWIN-TRANSITION-01)
Topic	HORIZON-CL4-2021-TWIN-TRANSITION-01-05 Manufacturing technologies for bio-based materials (Made in Europe Partnership) (RIA)
Consortium	17 organizations. 15 EU Member States + 2 non-EU state

GREEN LOOP CONSORTIUM PARTNERS

	Partner	Acronym	Country
1	IDENER RESEARCH & DEVELOPMENT	IDE	ES
2	NATIONAL INSTITUTE OF CHEMISTRY	NIC	SI
3	SLOVENIAN NATIONAL BUILDING AND CIVIL E. I.	ZAG	SI
4	FRAUNHOFER GESELLSCHAFT ZUR FOERDERUNG DER ANGEWANDTEN FORSCHUNG E.V	FHF	DE
5	LABRENTA SRL	LBRT	IT
6	MIXCYCLING SRL	MYX	IT
7	NERO SU BIANCO	NSB	IT
8	GERACE MARIA CRISTINA – TERRE DI ZOE'	TDZ	IT
9	IRIS TECHNOLOGY SOLUTIONS, SOCIEDAD LIMITADA	IRIS	ES
10	GLOWNY INSTYTUT GORNICTWA	GIG	PL
11	AACHEN UNIVERISTY: PROCESS CONTROL ENGINEERING / AACHEN UNIVERISTY: INSTITUTE OF SOCIOLOGY	AAU	DE
12	AUSTRIAN STANDARDS INTERNATIONAL	ASI	AT
13	INSTITUTO DE SOLDADURA E QUALIDADE	ISQ	PT
14	AXIA INNOVATION UG	AXIA	DE
15	ASOCIACIÓN DE INVESTIGACIÓN METALÚRGICA DEL NOROESTE	AIMEN	ES
16	NATIONAL COMPOSITE CENTER	NCC	UK
17	UNIVERSITY OF BRISTOL	UBRIS	UK

Funded by the European Union

Disclaimer: GREEN LOOP is a project funded by the European Commission under the Horizon Europe – HORIZON-CL4-2021-TWIN-TRANSITION-01-05- Manufacturing technologies for bio-based materials (Made in Europe Partnership) (RIA) under Grand Agreement Number 101057765.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or HADEA. Neither the European Union nor the granting authority can be held responsible for them.

Executive Summary

In recent years, the implementation of ML solutions and optimization techniques has significantly enhanced various aspects of manufacturing processes. By leveraging advanced algorithms and data-driven approaches, it is possible to predict and forecast different product-related aspects, that ultimately enables for a more efficient use of resources, and overall process productivity. In this regard, the present document gathers the work performed for the application of such methodologies to the materials considered in the project Green-Loop.

The results presented in this deliverable cover the different activities associated with the work performed during the completion of task 2.6 "Learning features towards optimization of material and products". The deliverable has been fulfilled by Idener and it provide a broad outline of the approaches taken to model the behaviour different bio-based materials submitted to specific manufacturing conditions including bio-rubber, bio-plastic, and wood composites.

For each material, a specific model has been created. A similar building methodology has been followed for each covering the entire model lifecycle. The following points describe the activities completed to support the development, validation, and implementation of these models:

- Problem definition, providing the basis and requirements that aim to develop the present work.
- Data acquisition, containing an overview of the datasets utilised, their content and other relevant data features.
- Data processing, describing the different cleaning and transformation techniques applied to data.
- Model selection, with and overview of the criteria applied to select the models used for each specific material according to the problem statement.
- Model training, covering the activities performed to train the models and the details obtained from such process.
- Model validation, showing the performance of the models generated according to specific metrics.
- Model deployment, describing how the models will be integrated into the complete Green-Loop platform.

TABLE OF CONTENTS

GREEN LOOP KEY FACTS 1

GREEN LOOP CONSORTIUM PARTNERS 1

EXECUTIVE SUMMARY 3

TABLE OF CONTENTS 4

LIST OF FIGURES 6

LIST OF TABLES 7

1 Introduction 9

2 Objectives 10

3 Literature review 11

4 Methods 13

4.1 Data preparation 15

5.2.1 Data cleaning 15

5.2.2 Feature engineering 15

5.2.3 Feature selection 16

5.2.4 Data splitting 16

5.2.5 Data normalization/standardization 17

4.2 Machine learning models 18

4.3 Model evaluation 18

5 Case studies 19

5.1 Bio-rubber 19

5.2.1 Bio-rubber: exploratory data analysis 19

5.2.2 Preprocessing bio-rubber data 25

5.2.3 Building bio-rubber model 26

5.2.4 Training bio-rubber model 28

5.2.5 Testing bio-rubber model 29

5.2 Bio-plastic 30

5.2.1 Bio-plastic: exploratory data analysis 30

5.2.2 Preprocessing bio-plastic data 34

5.2.3 Building bio-plastic model 35

5.2.4 Training bio-plastic model 37

5.2.5 Testing bio-plastic model 38

5.3 Wood composites 39
5.2.1 Wood composites: exploratory data analysis 39
5.2.2 Preprocessing bio-rubber data 44
5.2.3 Building wood composites model 48
5.2.4 Training wood composites model 50
5.2.5 Testing bio-rubber model 50
6 Prediction of temperature profile through convolutional neural network 51
6.1 Temperature profile from microwave heating: EDA 51
6.2 Preprocessing 53
6.3 Building model 54
6.4 Model testing 58
5.2.1 Transparent materials 58
5.2.2 Semi-transparent materials 62
5.2.3 Susceptor materials 70
7 Conclusions 76
8 References 77

LIST OF FIGURES

Figure 1. Stages of data science lifecycle 13
Figure 2. Tensile Strength data distribution 21
Figure 3. Tensile strength according to lignin and natural rubber content 22
Figure 4. Tensile strength according to material density and thickness of the test shield 23
Figure 5. Standardised dataset for bio-rubber data 26
Figure 6. Bio-rubber model training result 28
Figure 7. Bio-rubber model test prediction 29
Figure 8. Variation of elastic module according elements concentration 32
Figure 9. Bio-plastic elastic modulus by additive weights proportion 32
Figure 10. Elastic modulus element-wise scatter plot 33
Figure 11. Standardised dataset for bio-plastic data 35
Figure 12. Training result for bio-plastic: loss function and metrics 37
Figure 13. Bio-plastic model test predictions 38
Figure 14. Wood composites: number of samples for each material and type of test 42
Figure 15. Overall material behaviour for wood composites 42
Figure 16. Data cleaning representation for wood composites 44
Figure 17. Standardised data for V236 and tension test 46
Figure 18. Standardised data for V241 and tension test 46
Figure 19. Standardised data for V270 and tension test 46
Figure 20. Standardised data for V236 and compression test 47
Figure 21. Standardised data for V241 and compression test 47
Figure 22. Standardised data for V270 and compression test 47
Figure 23. KNN optimisation for compression 49
Figure 24. KNN optimisation for tension 49
Figure 25. Graphical description of microwave system trials 51
Figure 26. Convolutional layer representation 55
Figure 27. Max pooling layer representation 56
Figure 28. CNN top structure: transition from CNN to head block 56
Figure 29. Final model architecture 57
Figure 30. Boron nitrite: graphical testing results 58
Figure 31. Dense mullite graphical test results 60
Figure 32. Soda lime-glass graphical test results 62
Figure 33. Alumina cement graphical test results 64
Figure 34. Borosilicate glass graphical test results 66
Figure 35. Alumina silicate graphical test results 68
Figure 36. SiC graphical test results 70
Figure 37. ALN powder compact graphical test results 72
Figure 38. CuO graphical test results 74

List of Tables

Table 1. Bio-rubber variables..... 20
Table 2. Bio-rubber formulations..... 21
Table 3. Bio-rubber dataset description..... 24
Table 4. Inputs and output for bio-plastic case..... 25
Table 5. Hyperparameter tunning result for bio-rubber model..... 27
Table 6. Bio-rubber model test results..... 29
Table 7. Bio-plastic variables..... 30
Table 8. Bio-rubber formulations..... 31
Table 9. Bio-rubber, dataset description..... 34
Table 10. Inputs and output for bio-plastic case..... 34
Table 11. Search of optimal architecture for Bio-plastic model..... 36
Table 12. Bio-plastic model test results..... 38
Table 13. Wood composites, specimens used to perform laboratory tests..... 40
Table 14. Wood composites variables..... 40
Table 15. Wood composites dataset description..... 43
Table 16. Feature selection and role for wood composites..... 45
Table 17. Best KNN hyperparameter configuration for tension test..... 49
Table 18. Best KNN hyperparameter configuration for tension test..... 49
Table 19. Wood composites model test results..... 50
Table 20. Microwave temperature profile variables..... 52
Table 21. Microwave temperature materials..... 52
Table 22. Microwave heating dataset example..... 53
Table 23. Boron nitrate: numerical test results..... 59
Table 24. Boron nitrate: numerical test results..... 61
Table 25. Soda lime-glass numerical test results..... 63
Table 26. Alumina cement numerical test results..... 65
Table 27. Borosilicate glass numerical test results..... 67
Table 27. Alumina silicate numerical test results..... 69
Table 29. SiC numerical test results..... 71
Table 30. ALN powder compact numerical test results..... 73
Table 31. CuO numerical test results..... 75

ABBREVIATIONS

ML	Machine Learning
AI	Artificial Intelligence
DL	Deep Learning
IQR	Inter Quartile Range
EDA	Exploratory Data Analysis
ETL	Extraction Transformation and Load
ELO	Epoxidized Linseed Oil
DOP	Dioctyl Phthalate
TAG	Triacetylglycerin
KNN	K-Nearest Neighbours
MW	Microwave

1 Introduction

The growing demand for sustainable and environmentally friendly materials has led to a strong interest in bio-based materials. These materials are mainly obtained from renewable raw materials and are widely used in various industries due to their similar and in many cases superior properties to those of the original materials.

In this scope, the recent growing advancements in AI has paved the way for the development of ML techniques and statistical algorithms that present a valuable opportunity to further advance the understanding of the final bio-material properties. As example, the overcome of common problems faced by ML algorithms, and more specifically in the field of deep learning, has facilitated their reliability and generalisation as a possible modelling tool without inquiring on difficulties in training models. The most significant example of this advances is the work proposed by Xavier Glorot and Yoshua Bengio in [1], in which the problems of gradient vanishing and gradient exploding are solved, thus facilitating the training stage, the model convergence and demonstrating and enabling for the possibility of using deeper models, with the subsequent leaning capabilities.

The capabilities offered by ML can be leveraged in material science to accelerate the discovery and development of novel materials with specific properties. The strengths offered by this discipline surpasses traditional research methods, as usually they are time-consuming and expensive [2]. The accurate and quick predictions offered by these algorithms, set a new transformative paradigm in the material science by utilizing data-based solutions that mimic experimental results. In practice, these features are translated into speeding up the development of materials for their utilisation in heterogeneous domains including construction, industry, energy and storage and other technologies.

The advantages of using such kind of algorithms does not only rely on the prediction capabilities, as some software private frameworks allow for simulating specific material properties according to different processing conditions and feedstock features. Additionally, using ML approaches create an open-source environment that allows for a speeded-up computation compared to these traditional computing frameworks relying on complex mathematical computations.

Thus, this work integrates the ML techniques with the manufacturing of bio-materials to leverage the strengths of both disciplines, resulting in enhanced material properties and deeper understanding of the relationship between manufacturing conditions and final product features. It is expected that, ML can help to discover critical insights about the behaviour of the material through modelling specific material features. Later, this information can be used to optimise manufacturing processes or help to select the most appropriate materials for specific purposes.

2 Objectives

The main objective pursued under the course of the development of this work is to develop AI models that enable for optimizing the performance of novel bio-based materials.

Although the use of AI and ML is a mean to achieve such objective, the successful implementation of specific models that help to improve current solutions and that are tailored to the measurement and prediction of certain material properties is an extra objective added. Hence, this work does not only focus on generating value in the material science scope but also to leverage the most advanced data science and AI techniques and procedures to support it.

As mentioned, the work carried out touches different engineering fields, namely material science, AI, and data science. In the case of data science and AI, the quality and success of the project typically relies on the quality and amount of data available to find a solution that matches the expectations or that solves a given problem. Hence, the availability of data is a bottleneck that conditions the course of such kind of projects. In this regard, and with the aim of creating a solid data support, the activities carried out during this task are also oriented to collect and provide the larger amount of data of the highest possible quality, so data unavailability is not a limiting factor to fulfil the main objectives linked to this specific task.

Throughout the realization of this work, several key objectives have been pursued. One of the main challenges has been identifying the most appropriate ML algorithm suited to the characteristics of the data. While the specific material itself does not directly dictate the choice of algorithm, the nature and structure of the data do. The data represent the measured properties of the material, which vary in size, complexity, and quality. These factors ultimately influence the selection of an algorithm capable of making accurate predictions for the variables of interest.

Another important goal has been to ensure high performance across all materials under analysis. Configuring the ML approach to account for the unique properties of each material while maintaining consistency in model performance is crucial for delivering reliable results. This requires addressing the nuances and particularities of each dataset to achieve a well-rounded solution.

Finally, a critical objective has been the ability to package and deliver the solution in a reproducible format, ensuring that other companies can benefit from the methodology and models developed. Additionally, testing the solution to validate the created models is essential to ensure its robustness and effectiveness in real-world applications.

Funded by the European Union

3 Literature review

Before the rise of ML and DL, the prediction of material properties in material science using predictive models was based on combining empirical methods with theoretical models. These methods relied heavily on experimental design and simulation, which are highly time-consuming as they depend on accumulated experience. Currently, the combination of ML and material science is a growing trend in the research field, aiming to study and predict the properties of novel and advanced materials for their further use in various applications.

The most important aspect of developing a high-performance ML product is data. In general, data strongly determines the quality of the developed ML models, and it is influenced by both data quality and quantity. In the specific application of ML to material sciences, this data dependence is inherited from general ML principles and is highlighted in the studies reviewed [3]. In [4], data is identified as a critical factor for the success of an ML project, particularly for the application of ML in material sciences. Without sufficient data, ML models cannot capture insights or truly understand the implications and relationships present in the studied variables. In general, the suitability of applying ML to material science is well-supported, as the type of data typically used in material science is quantitative, and ML models demonstrate strong capabilities in this domain.

From the application of ML tools and methods in material science, a new field called materials informatics has emerged. This interdisciplinary field employs principles of informatics to enhance the understanding, utilization, and development of materials within the realm of material science [5]. The general approach to applying ML in material science involves collecting data, analysing and processing it, and constructing the model. Through this process, studies can gain insights, identify complex relationships, and discover trends in data, thereby linking material composition and manufacturing procedures to final material properties.

Within materials informatics, several subfields have been identified, focusing on different aspects of material science. According to [4], specific domains where ML is applied in material science include material property prediction, material development, microstructure identification, formation of multi-component alloys, and other areas. In alignment with the specific focus of the GREEN LOOP study, the application of ML in material science emphasizes the prediction of material properties, as well as hybrid classifications within the aforementioned research areas.

Material development focuses on how ML can predict specific material properties, such as mechanical, thermal, or optical properties. In [2], various ML models, including SVM, Decision Trees, and ANN, are used to predict the conductivity of new materials. The materials analysed include organic semiconductors for flexible electronics and perovskite materials across different compositions and processing conditions. The data for this study comes from various sources, including literature references, simulations using advanced computational systems, and experimental measurements of material conductivity. In general, the models demonstrated good performance, as they accurately reflected the structure-property relationships that govern conductivity in the analysed materials. In [6], a neural network is developed to predict ultimate tensile strength in aluminium alloys based on different chemical compositions and manufacturing processes. The variables considered include the content of aluminium, zinc, copper, or iron, and the metal treatments during manufacturing, such as annealing, strain hardening, or thermal treatment. The results highlight the importance of balancing model complexity with data complexity, and overall, good performances were achieved, with prediction errors ranging from 4% to 10% when estimating ultimate tensile strength.

Regarding the most suitable ML algorithms, there is no consensus on the most powerful model to use. Studies typically adapt and select algorithms based on the characteristics of the material science projects, the available data, and the specific objectives of the research.

The work presented in [7] supports the use of deep learning algorithms to suggest materials with specific properties by learning from databases that relate material composition, properties, and functionalities. The research focuses on materials where tailored performance is critical, optimizing the selection of materials that meet the desired properties and facilitating the design phase. This study considers mechanical and thermal properties, contributing to the creation of more durable, efficient, and sustainable materials. Other studies utilize complex multimodal neural network-based architectures to predict material properties [8]. In this case, the authors use chemical compositions and crystalline structures to predict conductivity, band gap, refractive index, and dielectric constant.

The application of ML to materials science is not limited to ANNs. Other classical ML algorithms are also effective in analysing and studying material properties. In some cases, algorithms other than ANN are more suitable and deliver better performance depending on the problem's characteristics. For instance, in [9], several ML models are tested to predict various mechanical properties of dental composites. The algorithms studied include XGBoost, AdaBoost, Random Forest, and KNN. To ensure an objective model evaluation, the authors used various performance metrics. The study supports the application of these algorithms for predicting flexural strength and Vickers hardness. Other studies combine ANN with classical ML algorithms. For example, [10] integrates KNN and ANN to build a model that imputes missing data in datasets used for predicting material properties.

The literature review emphasizes the application of ML methods for predicting material properties in materials science, noting that these approaches can surpass traditional methods based on empirical data and theoretical models. All studies highlight the critical importance of data in developing reliable models, where both data quality and quantity are essential for making accurate predictions. The application of ML in materials science is diverse, enabling the study of mechanical, electrical, and thermal properties, among others. There is no clear consensus on the best ML techniques to apply in this field. Rather, the selection of algorithms is determined by their suitability to the specific problem at hand.

4 Methods

The solutions proposed in this deliverable are data-driven approaches. Typically, data-driven approaches follow a sequential structure that helps data scientist to understand the world through data. The structure mentioned is also known as data science lifecycle and, in a broad and generalized way, can be illustrated as the figure above shows[11]:

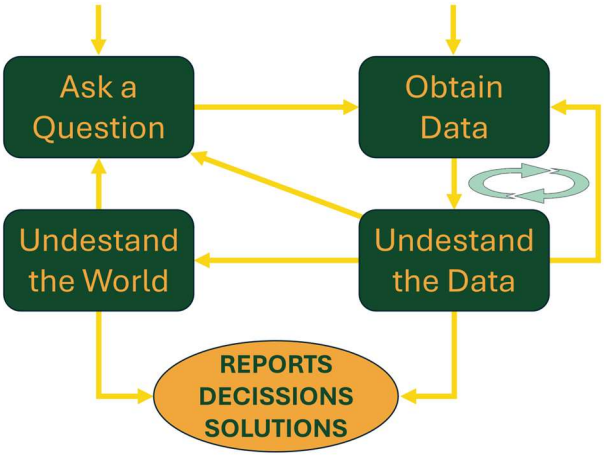


Figure 1. Stages of data science lifecycle

Ask a Question

The project starts with a necessity. A necessity arises when a question is raised. In this case, the question might be, how are the properties or behaviour of a specific material under specific manufacturing circumstances? What if the amount of additives is reduced? In this way, the necessity to understand in a descriptive way a certain real phenomenon is raised. This kind of questions helps to map the field in which the research is conducted, what kind of information would be useful and what patterns to detect.

In this work, the questions posed are driven by the concern of how the properties of a material respond to different shaping configurations. As the project deals with three different materials and each use case is unique, each use case will have an ad-hoc problem statement to specifically address the need to acquire knowledge about the material under analysis.

Obtain Data

Data forms the foundation for all decisions and conclusions in these types of projects. It serves as a representation of reality, allowing algorithms to capture and replicate real-world information. Establishing a robust strategy to collect high-quality, sufficient data is critical to ensuring the project's success. This stage is not only limited to collecting data but also also involves various data preparation activities, such as cleaning the data, performing feature selection, and applying other data transformation techniques to ensure the dataset is ready for analysis.

The datasets utilised in this work came mainly from work developed under the project GREEN-LOOP. Specifically, outcomes from WP3 - "Bio-rubber material production", WP4 - "Bio-plastic material production" and WP5 – "Wood composites material production" have been used to provide the basis on which later data science project stages will be sustained. Each WP has a couple of tasks dedicated to study the design, manufacture and technology validation of each material analysed. Both manufacturing conditions and

properties of the final products in the form of tabular data has been exported to be used during the execution of the T2.6.

Since the data from each use case will come from different sources, it's important to emphasize that preparing the data for the next step will require customized operations based on the specific needs of each dataset. The data may vary in terms of origin, structure, and format, necessitating different approaches for processing. For instance, data from one use case might be collected from physical experiments, while another could be sourced from simulations or sensor readings. As a result, the preparation process may involve handling varying levels of noise, missing values, or inconsistencies in the datasets which ultimately will require to develop a tailored ETL pipeline for obtaining data ready to be used to train ML models.

Operations made for preparing the data may include normalization, data cleaning, feature extraction, or even converting data into a uniform format. Tailoring the data preparation ensures that the subsequent analytical steps, such as modelling or pattern recognition, are based on accurate, clean, and relevant data that aligns with the requirements of each use case.

Understand the Data

In this phase, data scientists work with processed datasets to extract meaningful insights. These insights can be presented in form of visual information or through different statistical methods and AI algorithms that uncover key patterns and trends within the data.

As Figure 1 describes, this process is iterative, and it be submitted to several feed backs with the obtaining data stage. For example, if the insights gained are trivial or irrelevant, it may indicate the need for further data collection to enrich the dataset and include more valuable information that better captures the underlying phenomena.

Again, for the particular application of activities under this stage of the lifecycle, different methods might be required as the nature of the dataset is different in origin, shape and format. According to data requirements tailored data processing techniques will be applied. These techniques will depend on the type and quality of the variables contained in datasets and in the number of variables managed.

Understand the World

The former work will serve to answer the question raised at the beginning of the data science project. Both results and artifacts generated during the project can be used to better understand a specific parcel of knowledge area. In this case, the data science approach is sustained on the material science domain.

4.1 Data preparation

Data preparation procedures are commonly applied in data science projects. They are a way of guaranteeing that ML models learn from quality data, thus eliminating unnecessary and undesired entries or records and facilitating the ingestion of data to ML in an adequate structure, format and scale.

5.2.1 Data cleaning

The first step carried out in this stage is data cleaning. Here, raw data is analysed, and bad entries are detected and eliminated to remove noise and inconsistent data.

One of the main components of data cleaning consists of outlier detection or anomaly detection. Essentially, this stage consists of finding data objects with behaviours that are very different from expectation [12]. The reason behind applying this cleaning technique is that ML algorithms are highly sensitive to data quality [13]. For this reason, out-of-the-range values can introduce large errors that hinder the models training or they can provoke an overreaction during predictions and reduce models accuracy.

Using manual procedures is high time-demanding, and error-prone, so alternative methods have been applied to obtain the highest quality data. The alternative methods search for the automation in the procedure, so they are not specific for a certain material and are scalable when datasets increase in size with no additional effort. Specifically, for the detection of outliers inter-quartile-range (IQR) method has been used to remove values that are outside of the normal range. The IQR is the distance between the upper and the lower quartiles Q1 and Q3. Then the IQR range is applied as a filter considering the IQR n number of times, typically 1,5, and values outside this range are considered as outliers.

5.2.2 Feature engineering

Once the data has been cleaned, feature engineering is performed. This process involves creating new features that may provide better insights or enhance model performance. Common operations during feature engineering include extracting and transforming date components, encoding categorical variables, scaling or adapting variable ranges, and combining existing features to form new ones [14].

One frequently used transformation is the log transformation, which helps by compressing large values and expanding smaller ones. This process can often lead to obtaining a more symmetric, bell-shaped distribution, closer to normal, which is desirable for many modelling techniques.

Other typical techniques are aimed at reducing the dimensionality of datasets. These techniques apply linear transformations to create a new set of features with lower dimensions while preserving as much of the original information as possible. The most common technique is Principal Component Analysis (PCA), which identifies the directions (principal components) in which the data varies the most. PCA projects the original features into a new coordinate system where each axis represents a principal component, ordered by the amount of variance they capture from the data. By selecting only the top principal components, PCA effectively reduces the dataset's dimensionality, often improving computational efficiency and mitigating the risk of overfitting while retaining the essential patterns within the data.

5.2.3 Feature selection

With all the feature engineering done, its time to make a selection of the most interesting features from a modelling point of view. This stage identify the most useful features as the ones that provide the gross of the information to the model while at the same time simplifies the problem complexity and model size by removing redundant or irrelevant variables whose contribution to the final model performance is negligible.

5.2.4 Data splitting

After the feature selection stage, the next step is to split it into different subsets. Typically, three different subsets are created: training data, validation data, and test data. In supervised learning, each subset contains both features (input variables) and labels (target variables). It is desirable that each of the new subsets created the original distribution of the dataset is maintained. This will minimize the possibility of overfitting in later project stages. This aspect is especially relevant when dealing with unbalanced datasets where the presence of a specific class is very low compared to other classes present in the dataset.

- *Training Data*: this subset makes up the largest share of the original dataset. It is used to adjust the model's parameters during training through various optimization algorithms (e.g., gradient descent). The training data must be representative of the real-world problem to avoid introducing biases that can affect the model's performance. The quality and diversity of the training set directly influence the model's ability to learn relevant patterns.
- *Validation Data*: This subset is used to tune hyperparameters and prevent overfitting during the model training phase. Unlike training data, the validation data is not used to adjust model parameters; rather, it serves to monitor the model's performance and make decisions (e.g., early stopping, hyperparameter tuning) to improve generalization. If a significant difference between the performance on the training and validation datasets is observed (e.g., high training accuracy but low validation accuracy), this indicates overfitting. In such cases, adjustments are made to improve the model's generalization capabilities. In some data science project cross-validation technique is applied in place of a single validation set to make better use of available data and obtain optimal results.
- *Test Data*: this subset should never be used during the training or validation phases of the project. The test set is strictly for final evaluation, providing an unbiased assessment of the model's performance on unseen data. By withholding the test set until the very end, the final evaluation reflects the model's expected real-world performance. This approach prevents data leakage and ensures that performance metrics like accuracy, precision, recall, and mean absolute error are reliable indicators of how the model will behave in production.

5.2.5 Data normalization/standardization

Once the three different data subsets have been obtained, a standardisation or normalization step is carried out. This data transformation adjusts the properties of the dataset to formats that are more suitable for may machine learning algorithms. The main operations involved are data offsetting and range adaptation, which result in either data standardization or data normalization. These operations are particularly useful when variables included in the dataset are in different order of magnitudes or different ranges. It is demonstrated that ML algorithms perform better when all the features are on a similar scale[15]. The reason is that, through standardization or normalization, no single variable disproportionately influences the model performance.

- *Data standardization*: this process involves subtracting the mean of the data and dividing the result by the standard deviation. The application of this operation results into a dataset with mean zero and standard deviation 1. The formula for this mathematical operation is:

$$X_{standardized} = \frac{\mathrm{X} - \mu}{\sigma}$$

  Where:
  - $X_{standardized}$: dataset in standard format.
  - X: data in original format.
  - μ: mean of the dataset for each variable.
  - σ: standard deviation of the dataset for each variable.

- Data normalization: this operation scales the data to a fixed range, typically [0, 1]. For that, the minimum value from the observed dataset is subtracted for each variable. Then, the result is divided by the range, which is obtained by the difference between the maximum and minimum values. To perform data normalization, the following operation is carried out:

$$X_{normalized} = \frac{\mathrm{X} - X_{min}}{X_{max} - X_{min}}$$

  Where:
  - $X_{normalized}$: dataset normalized.
  - X: data in original format.
  - X_{min}: minimum value in the dataset for each variable.
  - X_{max}: maximum value in the dataset for each variable.

For this project, the technique applied will depend on the data being managed. The most suitable method to process the data will be applied to guarantee optimal results.

4.2 Machine learning models

ML models are an innovative way to reproduce real-world processes and phenomena by training algorithms through the utilisation of process related data.

There is a wide range of possible ML algorithms that vary in simplicity, capabilities and performance according to the task being managed. The purpose of algorithms inherently lies on data being processed: data is the material used to extract information about the underlying patterns, trends and insights that can drive predictive analytics. According to the dataset features, dataset size and other data characteristics, ML algorithms have been selected to achieve an optimal performance, finding the balance between use case suitability, computational efficiency and model accuracy.

Before selecting any algorithm, a rigorous study has been conducted to compare the performance of the different candidate ML algorithms.

Once a final model was selected, hyper parameter tunning has been applied to guarantee that model structure and configuration is optimal for its purpose. It is worth mention that each algorithm is customizable through various parameters and settings that make to optimise their performance and adapt them to specific problem requirements.

4.3 Model evaluation

According to the model purpose, different evaluation metrics can be used within the scope of ML. It possible to differentiate between the two main tasks performed by ML models: classification and regression.

The first one, classification aims to distinguish between a certain number of candidate classes by assigning a predicted probability to each. The closest the probability is to the target class, the better the model performance. According to the problem features, several metrics can be used in this scope, including accuracy, recall, precision and F-1, that are able to highlight the capabilities of the models for distinguishing various aspects of their predictive power, how they handle imbalanced data, or how they detect specific patterns in data.

On the other hand, regression models are used to predict a continuous variable. In this case, the metrics applied to assess model performance utilise the error between the prediction and the true value in different ways, including Mean Absolute Error (MAE), Mean Squared Error (MSE) or R-squared. These are robust metrics that provide a numerical and quantifiable measure of the accuracy and reliability of the models. In this way they can be used to compare models performance and thus select the model with highest prediction capabilities among a set of candidate algorithms or candidate model hyperparameter configurations.

5 Case studies

This section presents the technical application of the above-mentioned techniques to the use cases studied in Green-Loop. Three main materials are presented namely, bio-rubber, bio-plastic, and wood composites.

5.1 Bio-rubber

The first material presented in this deliverable is bio-rubber. As part of the GREEN-LOOP project, bio-rubber is being explored as a sustainable alternative for civil engineering and construction applications, thanks to its excellent fire-retardant and vibration-damping properties. These attributes make it a strong candidate for replacing conventional materials in critical infrastructure and construction projects, contributing to safer, more durable, and environmentally friendly structures.

The bio-rubber material is primarily composed of BASE_MATERIAL and is reinforced with lignin fibres to enhance its base properties. The use of lignin fibres is particularly advantageous due to their untapped potential. Despite being one of the most abundant natural fibres, only a small fraction is currently utilized, leaving significant room for improvement in fibre extraction and application processes. Lignin is a major byproduct of the pulp and paper industry, and leveraging it not only reduces waste but also contributes to more sustainable industrial practices. Furthermore, advancements in lignin extraction techniques could lead to further enhancements in the performance of bio-rubber, increasing its potential across various applications.

The following sections outline the methodology employed to apply AI in predicting the properties of bio-rubber. These are sustained by the application of ML techniques, and they provide an easy and powerful tool for material design and optimization. They are used as a predictive modelling tool and simulate material behaviour considering different material characteristics, allowing researchers and engineers to fine-tune the bio-rubber formulation to maximize its performance in real-world applications.

5.2.1 Bio-rubber: exploratory data analysis

Following the data science approach, the first step in building the bio-rubber model is to analyse the available data through an exploratory data analysis. This process enables a deeper understanding of the type of information being managed and its potential applications. The initial review of the dataset reveals the raw data with all the relevant variables. Overall, the data provided for this use case focuses on the physical and dimensional properties of the various bio-rubber compositions analysed. Specifically, it includes the name of the bio-rubber composition, the thickness (in millimetres) of the material, the lignin and natural rubber content (expressed as percentages), the density of the final material, and the measured tensile strength. The following table summarizes these results:

#	Feature	Description	Unit	Feature type
1	Name	Sample identifier for different compositions		Categorical
2	Thickness	The thickness of the bio-rubber sample.	mm	Numerical
3	Lignin loading	Proportion of lignin fibres added to the bio-rubber material	%	Numerical
4	Natural rubber	Amount of natural rubber in the material	%	Numerical
5	Density	The density of the final resulting material	g/cm^3	Numerical
6	Tensile Strength	The maximum stress the material can withstand while being pulled before breaking	MPa	Numerical

Table 1. Bio-rubber variables

The dataset structure consists of a set of interconnected variables that describe the overall tensile strength of bio-rubber materials. It includes various bio-rubber compositions characterized by different combinations of lignin and natural rubber content. In addition to compositional differences, the dataset also incorporates dimensional properties, such as shield thickness and density, which add further complexity to the analysis. For each unique combination of composition and dimensions, the corresponding tensile strength of the bio-rubber shield is recorded.

From this initial examination of the dataset, it is possible to identify a clear practical application of the information contained. Given a particular set of material requirements—such as tensile strength needed for a specific engineering or construction project—the dataset combined with ML techniques can be used to study the effects of different compositions and dimensions on material performance and help inform the design of new bio-rubber materials tailored for specific applications. This enables the optimization of bio-rubber formulations, ensuring that the material meets precise performance criteria while maintaining desirable environmental and sustainability characteristics.

The dataset comprises data collected from multiple laboratory experiments in which seven distinct species of bio-rubber were evaluated. Each species is characterized by varying concentrations of lignin and natural rubber, resulting in unique material properties. The dataset identifies each bio-rubber species with a unique alphanumeric identifier. Lignin concentrations range from 0% to 30%, while natural rubber content varies between 0% and 20%. These compounds are not combined in fixed proportions but follow an unspecified aggregation pattern, leading to variable compositions (e.g., low lignin and high natural rubber, or vice versa, with intermediate concentrations).

The following table shows the 7 different bio-rubbers analysed and the respective content in lignin and natural rubber:

#	Identifier	Lignin	Natural Rubber	Composition summary
1	NCC_GL_C_008	10	10	Moderate amount of both lignin and natural rubber
2	NCC_GL_C_009	0	0	No lignin or natural rubber present
3	NCC_GL_C_0010	10	0	Moderate lignin content, no natural rubber
4	NCC_GL_C_011	30	0	High lignin content, no natural rubber
5	NCC_GL_C_012	20	0	Significant lignin content, no natural rubber
6	NCC_GL_C_013	10	20	Moderate lignin content, high natural rubber content
7	NCC_GL_C_014	20	20	Significant amounts of both lignin and natural rubber

Table 2. Bio-rubber formulations

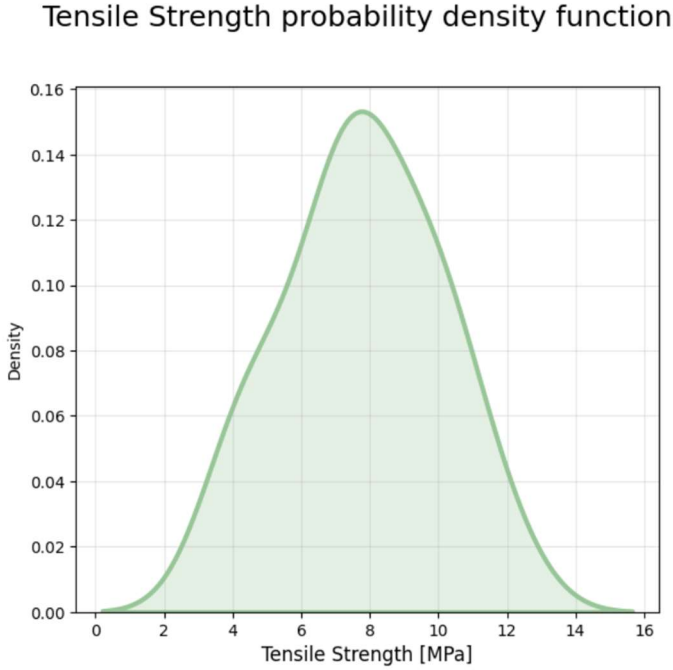


Figure 2. Tensile Strength data distribution

The property of the material object of study is the tensile strength. The different trials carried out yield a wide range of possible tensile strength for the produced and studied bio-rubbers. The above figure displays the probability density function for the observed data. For the different bio-rubber compositions, the essays provided the value in MPa, and they range from 4,05 to 11,79 MPa.

According to these results, the most probable tensile strength is 8MPa. Then, data shows a bell-shaped distribution, or quasi normal distribution, with a symmetric shape and moderate data spread indicating that values are clustered around the most frequent value. In this regard, a standard deviation of 2,23 MPa

indicates that material has consistent tensile strength, and that almost all the trials performed fall within a predictable range of 5,77 to 10,23 MPa.

As part of the exploratory data analysis, a scatter plot was generated to compare the influence of each compound on tensile strength. This visualization effectively illustrates the distinct impact patterns of both additives on the properties of the bio-rubber under analysis.

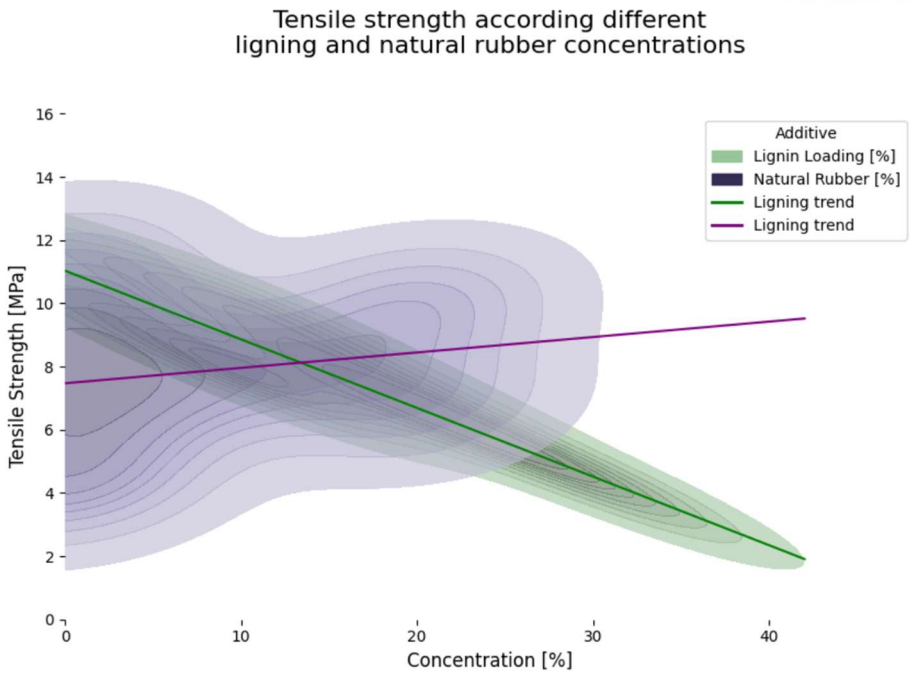


Figure 3. Tensile strength according to lignin and natural rubber content

For each material, different influences can be observed:

- Lignin loading: There is a clear decreasing trend with respect to tensile strength. As the lignin concentration increases, the tensile strength decreases. The green line represents the general variation or slope of the data. In this case, the data aligns closely with the line, indicating a clear linear relationship between lignin loading and tensile strength.
- Natural rubber content: The relationship between natural rubber content and tensile strength is less clear. A straight line has been drawn to represent the linear trend in the data, but the data is more dispersed, making it difficult to assume a linear relationship. The image suggests that the relationship between natural rubber concentration and tensile strength follows a non-linear trend, potentially influenced by other factors such as dimensional characteristics or the presence of other additives combined with natural rubber.

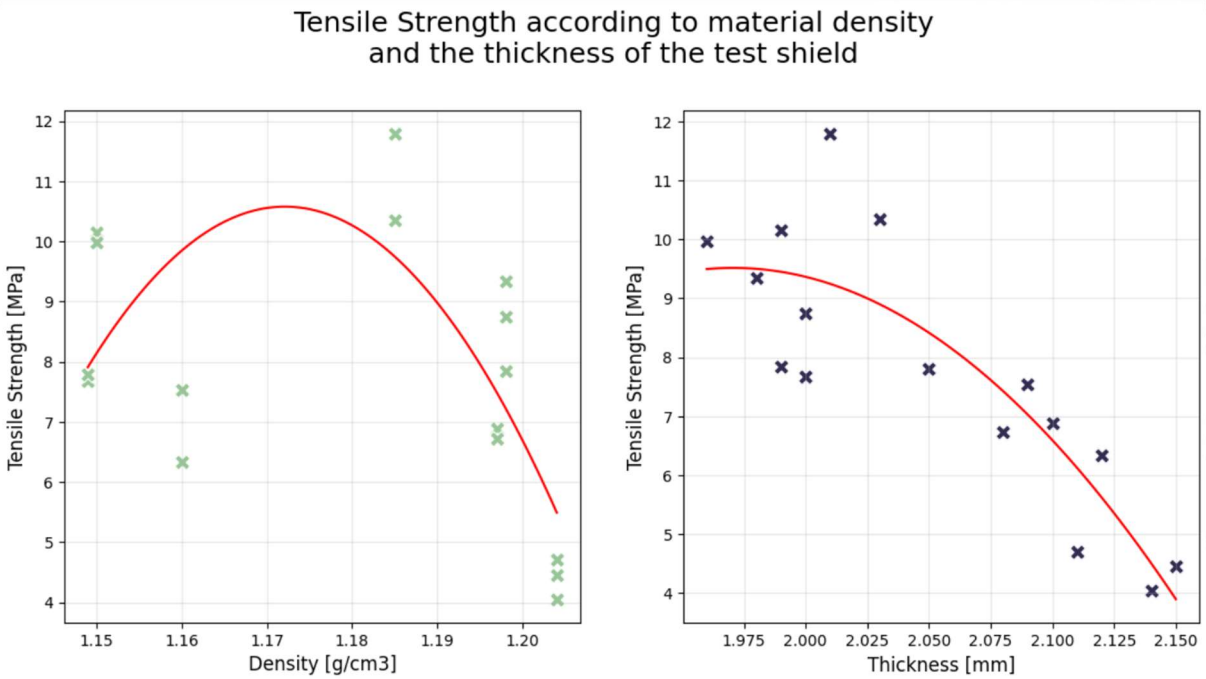


Figure 4. Tensile strength according to material density and thickness of the test shield

When analysing relationships between tensile strength and dimensional or physical characteristics of the test shield, some appreciations can be made. As Figure 4 shows, variable pairs indicate two opposite behaviours.

The left graph displays the relationship between the density of the test shields and the tensile strength they support for all the observations made during material testing. Densities measured range from 1.15 g/cm3 to 1.204 g/cm3 with a standard deviation of 0,023 g/cm3. In this sense, density values are quite stable across all the essays although within the given range, different appreciations can be made. Specifically, the relationship between both variables displays a concave parabolic trend, suggesting that tensile strength increases with density up to a certain point (around 1,175 g/cm3), after which it decreases as the density continues to rise and clearly displaying a non-linear behaviour.

When designing bio-rubber material with optimal tensile strength, the density curve indicates some valuable insights. Data suggests that maximum tensile strength corresponds to densities close to the peak of the curve, with a density of around 1,175 g/cm3. So, aiming for a material with maximum resistance, this density value should be obtained. As the density deviates from this peak value, resistance decreases. Reasons of this resistance variations are unknown, but they perhaps might be associated with structural factors such as increased brittleness or internal defects.

On right side graph, the relationship between shield thickness and tensile strength is showed. The thickness value ranges from 1,975 mm to 2,15 mm with a standard deviation of 0,062 mm. In this case, both variables display a non-linear quadratic relationship, although the trend is quasi linear.

Data displayed seems to be a semi-parabolic shape. This means that thicknesses below 1,975 mm would have been very useful in determining the actual relationship between both variables for the complete range of possible thicknesses.

According to data, a thickness slightly higher than 2 mm will deliver the best performance regarding resistance to tensile strength. Surprisingly, an increase in thickness do not improve the resistance of the

material, but the opposite effect, as data shows consistent decrease in tensile strength as thickness increases. According to these results, when designing bio-rubber materials with high tensile strength resistance, specific thicknesses should be taken into account to optimise this property of the material. Although out of the scope of this report, further investigation is recommended to analyse structural factors that determine the appearance of inconsistencies, or internal defects related to thicker shields.

This observation appears counterintuitive, as conventional reasoning would suggest that increased material thickness correlates with enhanced tensile strength. Regardless of material composition, this general assumption often overlooks underlying factors that may explain the inverse relationship observed. Understanding these factors could provide valuable insights for advancing bio-rubber material development. Incorporating detailed manufacturing parameters into the dataset or conducting a thorough analysis of the production process could help identify sources of variability, reveal potential defects, and guide optimization strategies to reverse this unexpected trend.

Finally, the following table gathers the description of variables considered in this study from a statistical perspective. Different descriptors are included to better understand the ranges in which each variable is presented, and the distribution of each alongside the experiments.

	Thickness [mm]	Lignin Loading [%]	Natural Rubber [%]	Density [g/cm3]	Tensile Strength [MPa]
mean	2.05000	15.000000	6.250000	1.180500	7.774374
std	0.06261	9.660918	8.850612	0.022871	2.236136
min	1.9600	0.0	0.0	1.1490	4.0531
25%	1.9975	10.0	0.0	1.1575	6.6300
50%	2.0400	10.0	0.0	1.1910	7.7379
75%	2.1025	20.0	12.5	1.1980	9.5012
max	2.1500	30.0	20.0	1.2040	11.7969

Table 3. Bio-rubber dataset description

5.2.2 Preprocessing bio-rubber data

Once exploratory data analysis indicates the possibilities when working with the provided dataset, data is processed to obtain a good quality dataset to be feed into the model. Data raw contains the variables indicated in table Table 1. All these variables have been processed according to their characteristics, and role within the model and later use of the model.

The first step was to clean the dataset by removing the bad entries or noise entries present in the dataset. In this way, the model will not focus on out-of-range values, and it will be centred on the normal or actual operation range.

A feature selection process was then conducted to evaluate the relevance and potential impact of each variable in the model-building process. From the initial list of variables, the test identifier or name was discarded, as it provides no material-related information and potentially adds noise to the model, leading to undesirable results. Moreover, any potential information conveyed by this variable would be redundant, as both lignin load and natural rubber content are directly related to it. Instead of removing key compositional features, eliminating the identifier is a more effective approach. This is because naming conventions may change over time, introducing inconsistencies, while variables such as lignin content and natural rubber content are more stable, measurable, and directly tied to the performance of the material.

Then the final list of variables utilized by the model are:

Variable	Type
Thickness	Input / feature
Lignin loading	Input / feature
Natural Rubber	Input / feature
Density	Input / feature
Tensile Strength	Output / target

Table 4. Inputs and output for bio-plastic case

The final data preprocessing step before model deployment is data standardization, which will be applied to both the input and output variables. Standardization adjusts each variable to have a mean of zero and a standard deviation of one, which enhances the training process, improves model performance, and strengthens generalization by ensuring all features contribute equally to the learning process.

Separate standardization models have been built for the input and output variables. This is because each preprocessing model performs a specific role during production or validation. When a new data point is received, the input preprocessing model will standardize the raw data, converting it into a form suitable for

the model. After the model generates predictions, the output standardization model will reverse the process, transforming the standardized outputs back into their original scale.

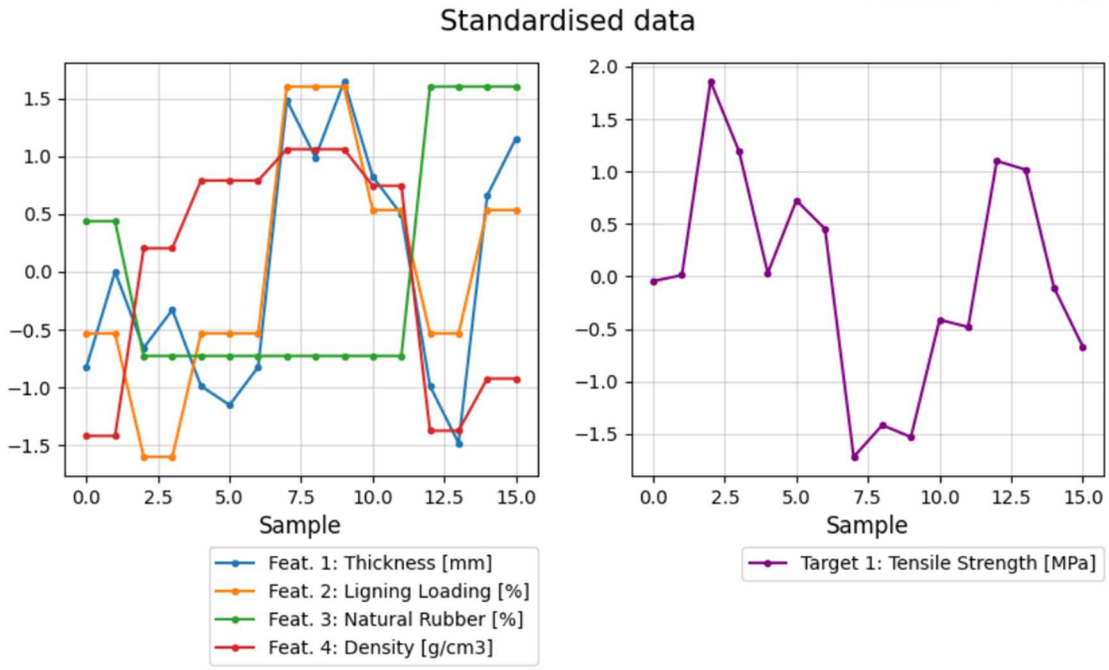


Figure 5. Standardised dataset for bio-rubber data

5.2.3 Building bio-rubber model

The building of the bio-rubber model relied on the design of a neural network that reliably reproduced the patterns found in data. The determination to use this algorithm has to do with the potential, flexibility and ability to model non-linearities, as EDA discovered on data.

The next step when modelling the bio-rubber behaviour is to find the optimal model configuration and architecture. Although manual methods can return acceptable results, automating the labour of finding the optimal model architecture ensures trying a higher variety of model configurations and assess the performance of such models during a low range of training epochs. For this process, Keras-tuner[16] library was used to facilitate the search of optimal hyperparameters. This framework allows for creating a hypermodel that is basically a Keras[17] model structure that contains all potential model configurations to be tried. Different optimisation configuration can be set, including the optimization mode, the number of trials per run, and the number of epochs performed in each run, so a balance between time and performance can be obtained by adjusted both parameters. After performing a set of trials, the framework returns the summary of the exploration, so they can be used to rank all the model architectures tried and sort models according to a specific performance metric.

For this case, three different performance metrics where assessed: MAE, MSE and R2. Although MAE and MSE provide a general idea about how model performs, R2 provides a deeper understanding of how model behaves, so the metric used to select the model was R2.

According to data type managed, model building utilised fully connected layers as they are suitable for processing structured data and model non-linear real-world phenomena.

In addition to model architecture, which basically relies on finding number of layers and number of neurons per layer, other model hyperparameters were optimised. Specifically, the type of activation function for each potential layer added to the model and the optimal learning rate to train the model. For the case of activation functions, linear, sigmoid, ReLU and tanh were considered in the search. Although the potential benefit of these hyperparameters is marginal, every possible improvement made on model building is positive to create a good final result.

In this case, optimisation was performed in two stages. The first stage was focused on finding architecture-related hyperparameters such as number of layers and number of neurons per layer with a wide search space. Although not a definitive result, this first exploration sets a baseline for a further model optimisation, which has to explore around the results obtained during the first trial. Accordingly, the second model optimisation stage was made around the first result obtained, narrowing the searching space but adding a higher resolution on the searching space. Optimal model configuration was extracted from this later optimisation which included the optimal model structure with the number of layers, number of neurons per layer, type of activation function used in each layer and the optimal learning rate.

Both optimisation processes considered only two constraints. The first one consisted in limiting the number of neurons that layer n+1 has according to the number of neurons use for the layer n. This approach aimed at finding a model structure that refine the information according to it flows through the network. In this sense, a pyramidal architecture is prone to be obtained. Otherwise, hyperparameter optimisation techniques might rely on large searching spaces, which would generate variable network width, with continuous narrowing and shrinking in the number of layers that do not obey to any technical or performance criteria, only randomized factors due to the large size of the search space. The second constraint is related to the input and output layers. Since input and output tensor shapes are known, these layers must have a specific shape to be able to process tensors of the expected dimensions.

A total of 300 trials were performed in each optimisation stage. Models were trained for 5 epoch in each trial, which is enough to allow metrics evolve and decide the model performance accordingly. After this process, the results obtained and model configuration is presented:

Parameter	Optimal configuration
Number of hidden layers	3
Units layer 1	45
Activation layer 1	tanh
Units layer 2	26
Activation layer 2	ReLU
Units layer 3	6
Activation layer 3	ReLU
Units output layer	1
Activation function output layer	Tanh
Learning rate	0,00484

Table 5. Hyperparameter tunning result for bio-rubber model

5.2.4 Training bio-rubber model

The model architecture described earlier was constructed and saved for use in subsequent stages of the project. The initial step in training involves loading the model. During this phase, it is crucial to consider the ancillary artifacts associated with the model. In this case, the model is not solely the neural network but also includes an object to evaluate model performance, specifically the R^2 (coefficient of determination) function.

Once the model is loaded and instantiated, it is compiled—a process that prepares the model for training by setting key parameters: the optimizer, the loss function, and the performance metrics. The optimizer, loss function, and metrics guide the training process, help in evaluating the model's performance, and prevent overfitting. This is further complemented by various callbacks, such as early stopping, which terminates training when overfitting is detected. Overfitting is usually identified through discrepancies in performance between the training and validation datasets.

The training configuration utilized includes the ADAM (Adaptive Moment Estimation) optimizer with a learning rate optimized in earlier sections, the Mean Absolute Error (MAE) as the loss function, and MAE, Mean Squared Error (MSE), and R^2 as performance metrics. This setup allows us to track the model's performance across each epoch and evaluate differences between the training and validation subsets. Notably, the MAE serves both as a metric and the loss function, guiding the optimization algorithm during model fitting.

The model was trained over 100 epochs with a batch size of 8 samples. The minimal variation in performance metrics between the training and validation sets indicates robust generalization capabilities, with favourable outcomes across all metrics (low MAE and MSE, and an R^2 close to one). These results reflect the effectiveness of prior optimization steps, including refining the model architecture and optimising the model hyperparameters.

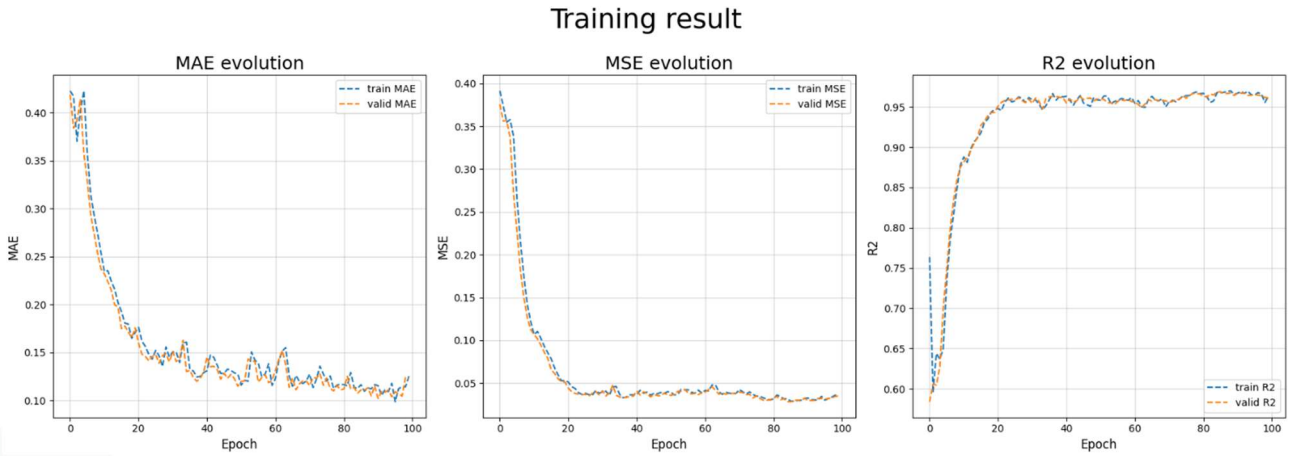


Figure 6. Bio-rubber model training result

5.2.5 Testing bio-rubber model

The final stage in developing the ML-based model for predicting the tensile strength of bio-rubber materials, based on structural and compositional features, is model testing. This testing phase integrates the neural network with the standardization models discussed in previous sections, enabling a more accurate assessment of the prediction deviations in real-world magnitudes.

To evaluate the performance of the model, the MAE, MSE, and R^2 metrics were employed. MAE was used to measure the average magnitude of errors without bias towards the size of individual errors, providing a straightforward interpretation of prediction accuracy. MSE, on the other hand, was utilized to identify the susceptibility of the model to large errors due to its tendency to penalize larger deviations more heavily. Finally, R^2 was used to evaluate the proportion of variance in tensile strength that is explained by the model, serving as a key indicator of overall predictive performance.

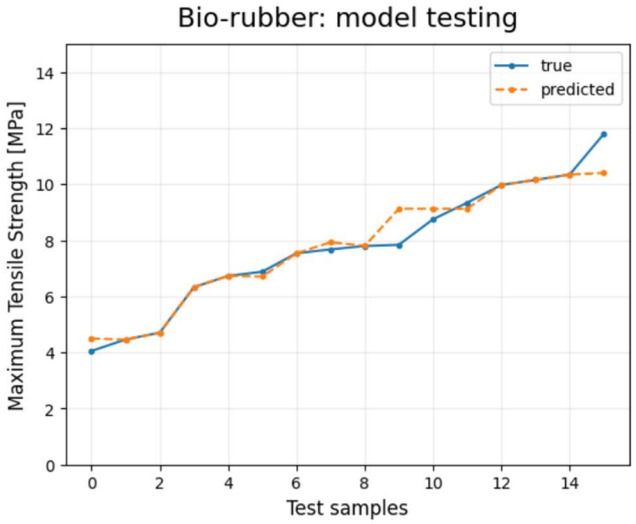


Figure 7. Bio-rubber model test prediction

Model	MAE [MPa]	MSE [MPa^2]	R^2
Optimised ANN architecture	0,2650	0,1640	0,9659

Table 6. Bio-rubber model test results

Overall, results are indicative of a high performance. With an average tensile strength value of 7,77 MPa, and average error of 0,265 MPa is quite accurate, considering that in relative figures this amounts for a 3,41 %. Also, the model does not commit large errors in predictions, as the MSE expresses. Finally, the high R2 value obtained, very close to 1, indicates that the model effectively captures the variance in data. In summary, the model demonstrates strong predictive power and reliability in estimating tensile strength based on the given input features.

Once model was tested internally by IDENER it was validated externally by NCC. For that, model was enveloped in an executable that contains all the required software requirements and artifacts, so the external validation process was straightforward. NCC validated de model by performing some specific tests, and they could confirm the results obtained by IDENER.

5.2 Bio-plastic

Bio-plastic materials are the second focus of study within the GREEN LOOP project. These materials represent a promising alternative to conventional plastics across a wide range of industries, including packaging, household appliances, electronics, and agricultural products. The key advantage of bio-plastics lies in their ability to offer mechanical and technical properties comparable to those of traditional plastics, while also introducing the significant environmental benefit of biodegradability. This makes them an invaluable option for replacing older, less sustainable materials. Recent advances in biotechnology and manufacturing techniques have significantly boosted the development and application of bio-plastics, bringing them closer to mainstream adoption. Notably, the ability to incorporate waste from other sectors as raw material enhances their appeal, offering a clear path toward reducing environmental impact and contributing to circular economy practices, where materials are reused and recycled, leading to a more sustainable product lifecycle.

In the context of Task T2.6, the work performed delves into the intersection of bio-plastics and AI-driven methodologies. These techniques were employed to predict the properties of bio-plastic materials with a high degree of accuracy, allowing for improved material design and performance optimization. The following sections outline the specific approaches taken to harness the power of machine learning in modelling and forecasting key bio-plastic characteristics, demonstrating how data-driven insights can facilitate the broader adoption of these eco-friendly alternatives in various sectors.

5.2.1 Bio-plastic: exploratory data analysis

The dataset utilized for the building of the model representing the bio-plastic material contains different features, which each of them contains different fields of information. Features include the formulation of the material composition, the percentage in weight of additives, the carbon content proportion, the hydrogen content proportion, the oxygen content proportion, and other features containing properties of the material as the elastic module.

#	Feature	Description	Unit	Feature type
1	Formulation	The composition of the material mixture		Categorical
2	% weight	Weight percentage of the additive in the formulation		Numerical
3	C	Carbon content proportion in the mixture		Numerical
4	H	Hydrogen content proportion in the mixture		Numerical
5	O	Oxygen content proportion in the mixture		Numerical
6	Elastic modulus	Calculated elastic modulus of the material	MPa	Numerical

Table 7. Bio-plastic variables

The dataset presented provides an overview of different bio-plastic formulations. The formula or material characterisation is provided by the feature "Formulation", and it expresses different formulations of polyhydroxybutyrate (PHB) combined with various plasticizers or additives.

The key concept here is that the different formulations given by the addition of plasticizing agents modify the properties of the base PHB. The variation on the final properties obtained are measured through the elastic module of the material. According to this property, it will be possible to later classify the obtained material to specific applications according to its suitability and expected performance.

The additive content value is provided in two different formats: phr (parts per hundred of rubber) that means parts of additive per hundred rubber and wt% (weight percentage), that refers to weight percentage of the additive in the formulation. Although the formula is expressed using these units, the final numerical ratio of additives is always expressed as weight percentage.

Specifically, 6 different formulations are analysed:

#	Formulation	Additive	Description
1	PHB + 15 phr ELO	Epoxidized Linseed Oil	This formulation expresses that for every 100 parts of PHB 15 parts of ELO are added. Typically, ELO is used to improve flexibility and plasticity in the polymer.
2	PHB + 5 phr ELO	Epoxidized Linseed Oil	This formulation has a lower amount of ELO. It exactly contains 5 parts of ELO per hundred of PHB. This low concentration of ELO will probably result in a less flexible material compared to the previous formulation.
3	PHB + 30 wt% DOP	Dioctyl Phthalate	The base content of PHB is complemented with 30% by weight of DOP. This additive is used to increase the flexibility of the base biopolymer.
4	PHB + 10 wt% DOA	Dioctyl Adipate	To the base PHB, 10% in weight of DOA is added. This additive helps to plasticize the PHB and thus it increases the flexibility and durability of the final polymer.
5	PHB + 20 wt% TAG	Tryacetylglycerin	This formula includes 20% by weight of TAG. TAG is used as a plasticizer as it improves the flexibility and processability of PHB, potentially enhancing its impact resistance and flexibility.
6	PHB + 10 phr ESBO	Epoxidized Soybean Oil	This formulation introduces a new material, ESBO. In this case, the sample contains 10 parts of ESBO per hundred rubber of PHB. This additive is a plasticizer and stabilizer whose properties help to improve the thermal stability and flexibility of the final compound.

Table 8. Bio-rubber formulations

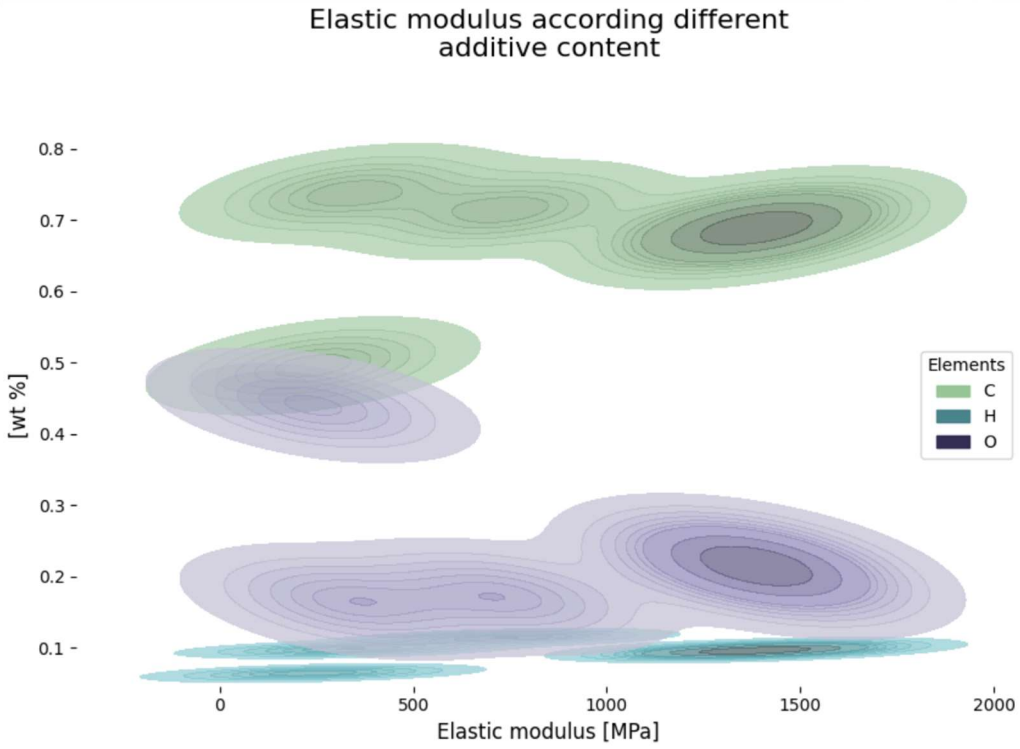


Figure 8. Variation of elastic module according elements concentration

As observed, the basic elements in the final polymer have varying proportions in its overall composition. The most abundant element in the molecular structure of the polymer is carbon, followed by oxygen, and lastly, hydrogen. Although there is a noticeable spread in the distribution of the elastic modulus values, certain trends can be identified as the composition of each material changes. These trends, however, are not consistent across all elements. Depending on the specific element being analysed, the elastic modulus may increase or decrease with the addition of a particular additive containing that element.

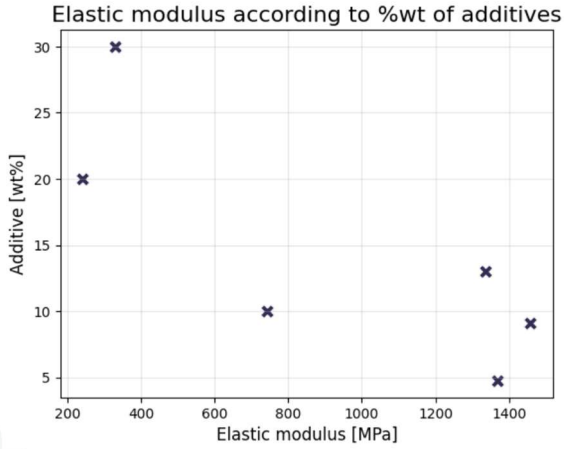


Figure 9. Bio-plastic elastic modulus by additive weights proportion

The figure shows the elastic module value according to the percentage in weight of additives added. Using this variable does not provide perfectly clear results. Although the scattered data points suggest a potential increase in the elastic modulus with small amounts of additives, the limited number of data points combined

with the relatively high variability at similar additive concentrations makes it difficult to draw definitive conclusions about the additives' influence on the final polymer properties. This ambiguity arises because each concentration value corresponds to a specific additive, each with unique properties that ultimately affect the polymer's behaviour in different ways.

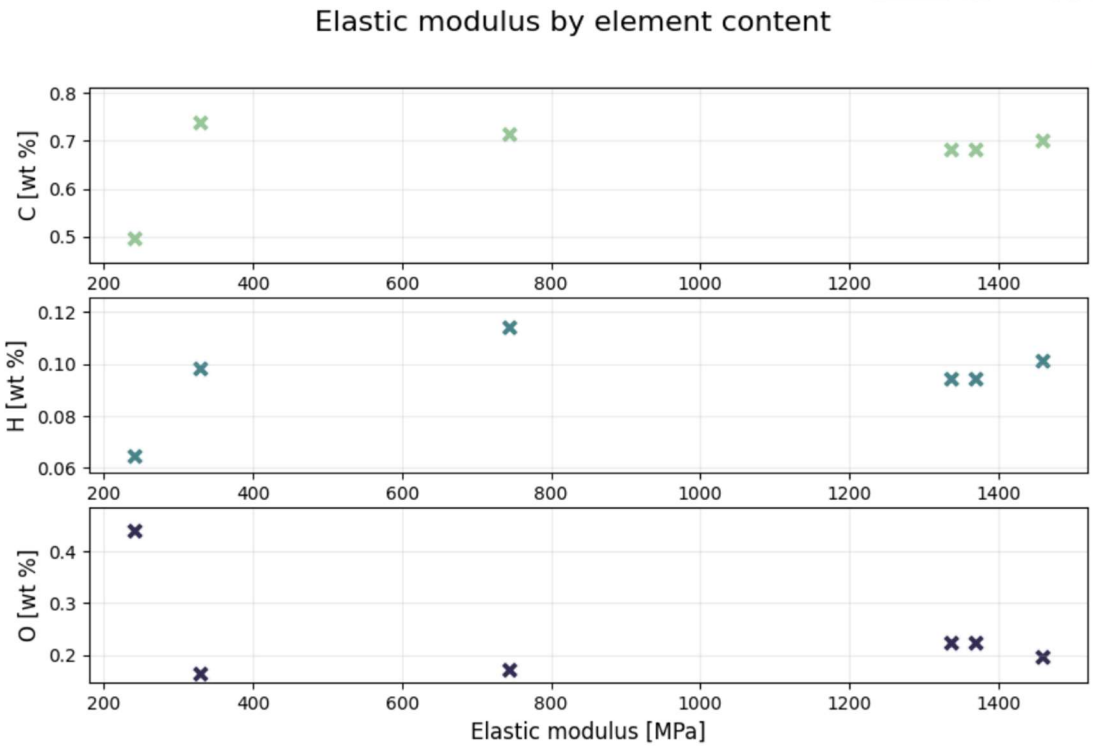


Figure 10. Elastic modulus element-wise scatter plot

When analysing the data element by element, attempting to correlate the presence of specific elements in the polymer formulation with its elastic modulus, no clear trends emerge. This suggests that the relationship between the elemental composition and the elasticity of the material is more intricate than a simple linear correlation. In fact, no linear patterns or gradual changes in properties are observed. This indicates that the underlying relationship between the composition of the polymer and its elasticity is likely non-linear and more complex. A deeper investigation into the interactions between elements and the overall material structure is required to fully understand how the composition of the polymer influences its mechanical properties.

	wt %	C	H	O	Elastic modulus [MPa]
mean	14,48	0,669	0,094	0,236	913,085
std	9,132	0,087	0,016	0,102	549,078
min	4,762	0,495	0,064	0,163	241,228
25%	9,318	0,682	0,094	0,178	432,347
50%	11,521	0,692	0,0962	9,210	1040,249
75%	18,260	0,710	0,100	9,223	1361,417

	wt %	C	H	O	Elastic modulus [MPa]
max	30,000	0,738	0,114	0,440	1458,596

Table 9. Bio-rubber, dataset description

5.2.2 Preprocessing bio-plastic data

This section describes the sequence of steps performed to prepare and process the bio-plastic data to create a ML model. It should be mentioned that this process required specific and tailored data transformations according to the characteristics of the dataset utilized for this use case. At the end of this stage, a dataset in optimal conditions to train a ML model will be obtained.

The use case started from a file containing a set of records for the variables described in Table 7. Bio-plastic variables. No bad entries were detected during the data cleaning stage, so no need for input wrong values or missing values.

The feature selection stage discarded one of the variables included in the original dataset. In this case, "formulation" variable was removed from the final list of features used to build the model. The reason is that the variable contained metadata about the additive used to be combined with the base polymer rather than numerical information interpretable for mathematical models. After this feature selection, the final list of variables was:

Variable	Type
wt %	Input / feature
C	Input / feature
H	Input / feature
O	Input / feature
Elastic modulus	Output / target

Table 10. Inputs and output for bio-plastic case

Given the format in which the data was presented, no additional feature engineering was necessary. All variables were already in a numerical format and were evenly distributed, although they had different value ranges and orders of magnitude. Therefore, standardization was applied to scale all the variables to a common range.

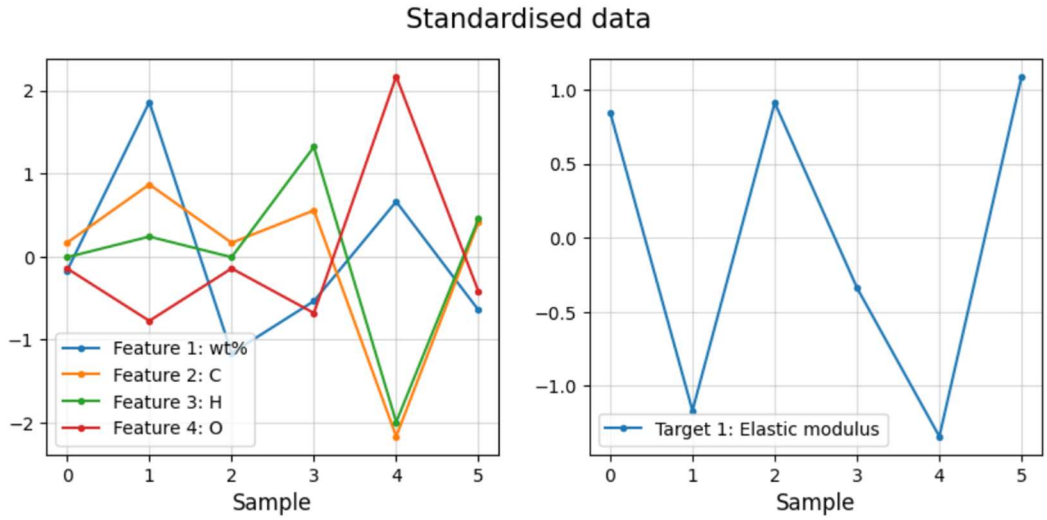


Figure 11. Standardised dataset for bio-plastic data

The previous image shows the data aspect after standardisation. As is it possible to observe, all data values are around zero and the have a similar domain. This is the exact result desired when performing data standardisation. Now data is ready to be used to train a ML model for a regression task. The model will be in charge of mapping the relationship between the data in the left image and the data in the right image.

5.2.3 Building bio-plastic model

This section describes the work behind building the ML model for the bio-plastic material. For this use case, the algorithm selected was neural networks according to the versatility and capabilities when modelling complex and non-linear systems. The basic structure of such networks consists of a series of sequential stacked layers that refine the information passing through to ultimately obtaining the final predicted value. In this case, fully-connected layers were employed, as no specific layer typologies were required given the characteristics of the data utilized.

The process of building machine learning models based on neural networks is not an exact science. Different approaches can lead to acceptable solutions with similar results. As such, complexity is not always directly related to performance. This means that it is possible to create simple models with higher prediction accuracy and better generalization capabilities than more complex models with a larger number of parameters.

For this reason, different model architectures were tried to find the optimal balance between complexity and performance. Models built range from different complexities both in number of layers as in the number of neurons per layer, resulting in models with different number of parameters. Table 11. Search of optimal architecture for Bio-plastic model summarises the different models tried and the obtained performance, as well as some details about the architecture of each.

They all of them share a common structure for the output layer. Considering that the model has to predict a single variable (Elastic modulus) the last layer will be essentially composed of a single neuron. It is crucial to configure this final neuron according to the transformations applied during data preprocessing. Specifically, the activation function must align with the range of the target variable. In this case, since the target variable values slightly exceed the range [-1, 1], using an activation function with bounds narrower than this range would limit the performance of the model. The model would be unable to output values beyond the limits of the activation function, potentially resulting in inaccurate predictions. For this reason, the last layer uses

linear activation function as it does not have any limitation in the values that is able to yield unlike other commonly used activation functions as ReLU, sigmoid, or hyperbolic tangent.

Model	Layers	Neurons per layer	Activation functions	Params	Valid. MAE	Valid. MSE	R^2
1	6	256|32|…|32|1	ReLU|…|ReLU|Linear	12705	0.934	0.9683	0.0317
2	5	256|32|32|32|1	ReLU|…|ReLU|Linear	11649	0.964	1.0427	-0.0426
3	4	256|32|32|1	ReLU|…|ReLU|Linear	10593	0.989	1.1188	-0.1187
4	3	256|32|1	ReLU|…|ReLU|Linear	9537	0.030	0.0023	0.9977
5	3	20|10|1	ReLU|…|ReLU|Linear	321	0.018	0.0012	0.9987
6	2	20|1	ReLU|Linear	121	0.016	0.0010	0.9989

Table 11. Search of optimal architecture for Bio-plastic model

As it is possible to observe, the model performance does not obey to size reasons, but to suitability of the architecture to the problem being solved. There is a substantial shift in model performance when the number of layers decreases below 4. Although the number of model parameters in this case is not significantly lower than the model with one more layer, its overall complexity does. This means that the data insights do not require such large amount of neurons interconnections, and that this additional complexity generates noise in the predictions that are unbeneficial for the model performance. Just though this change, the model improves its performance in almost two orders of magnitude considering the score obtained in MAE. Similarly, for the case of MSE, score improvement are propagated in twice in magnitude orders. Generally speaking, the model is more accurate in its predictions and when it commits a mistake, errors are not so far from the true value. In this regard, the most clarifier metric is the R^2; only when model complexity decreases below four layers the model starts predicting properly. For model 1, the R^2 value indicates a bad performance considering that the value obtained is similar to always predicting the mean value of the predicted variable. It is even worst for models 2 and 3 since a negative value indicates that the model performance is worst than always predicting the mean value of the elastic modulus.

Moving on simplest models, (model 4, model 5 and model6), it is appreciable a steady improvement in model scores. These improvements are consistent with the reduction of the number of parameters composing the model, although the overall difference in performance of the models is almost negligible. Values obtained for R^2 are clear evidence that the model is able to predict properly. Moving on in the problem solving, the architecture with the best scores with validation data is selected to create the final model.

According to the results shown, the best model architecture is the architecture of the model 5. This model is composed of two layers: the first one with 20 neurons and the output layer with one neuron. The input layer has ReLU activation to be able to reproduce non-linearities present in the material behaviour. The output layer has linear activation function to avoid limitations in the output value.

5.2.4 Training bio-plastic model

To train the selected model architecture, different considerations and training configurations were made. Firstly, model overfitting was avoided by including training callbacks which monitor a specific metric to perform early stopping and avoid model overfitting the training data. In this case, the callback tracks the MAPE to stop the training stage as soon as a specific value of MAPE is detected. This allowed for setting a high number of training epochs, as the algorithm is able to stop the training by itself in optimal conditions.

The model training configuration was set by defining the optimizer, loss function, and monitoring metrics. For the optimizer, in charge of performing model weight updates in each epoch, the Adam algorithm was used. For the loss function, which guides the optimization algorithm during the training stage, MAE was selected. Finally, MSE and MAPE were selected as metrics to track and monitor the performance of the model during the training and compare these scores both with the training data and the validation set.

Now, training results are presented. Both loss function and MSE metric evolution are displayed to assess how the model evolves across each epoch of the training stage.

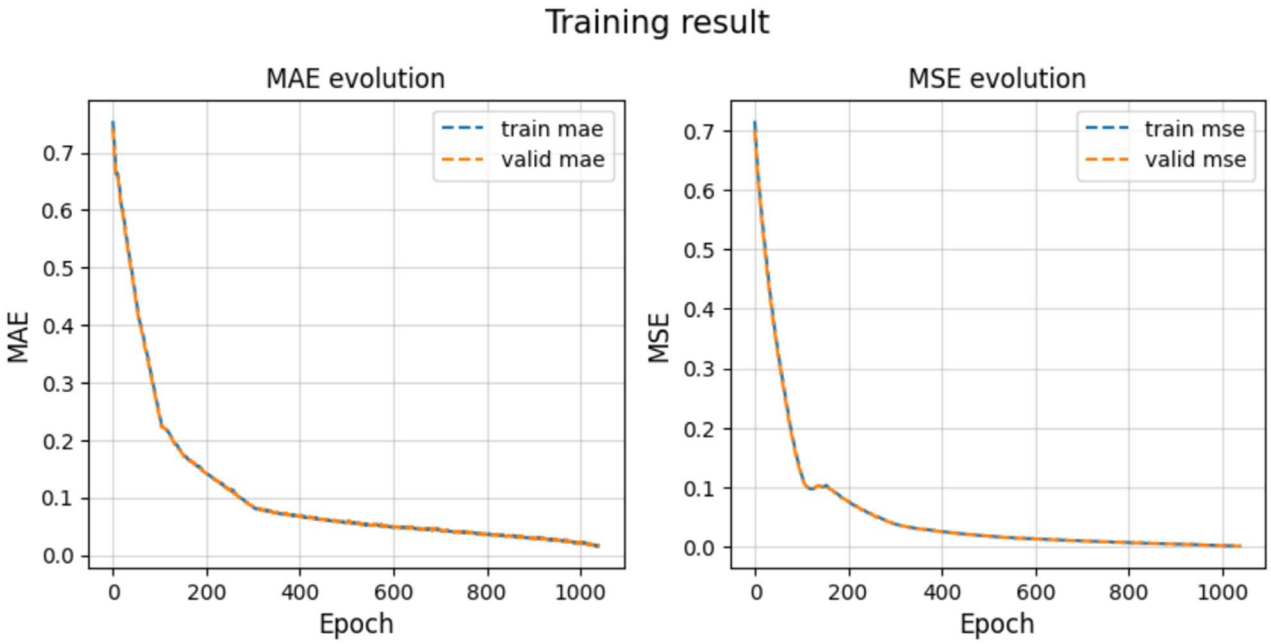


Figure 12. Training result for bio-plastic: loss function and metrics

As the graphs show, both the loss function and the metrics continuously decrease until they reach a steady value. During the initial epochs, the slope of both functions is steeper, indicating that in each epoch, the model improves significantly. After 100 epochs, the slope of the loss function starts decreasing until it reaches a steady decrease, which continues until the end of the training. In both plots, the model performs adequately, as the differences between training and validation are negligible. The training stops when the callback detects no meaningful improvements.

Finally, it is important to highlight the difference between the two curves. This is because MSE penalizes large errors more, while MAE treats all errors equally, regardless of their magnitude. This is why the MSE curve has steeper slopes—initially, even minor improvements significantly reduce the error. However, after the elbow, many epochs are required to achieve further significant error reduction.

5.2.5 Testing bio-plastic model

For testing the bio-plastic model, test data subset has been used. For testing the model both MAE and MSE have been used, MAE to measure the performance the average error value and MSE to assess if the model commits large errors. Unlike previous sections, the model performance will be assesses using values in real magnitude, i.e, expressing the errors as elastic modulus values. In this way, is possible to evaluate the model in a potential production environment considering that it will receive requests to obtain specific properties of a bio-based material according to certain material composition and additives used.

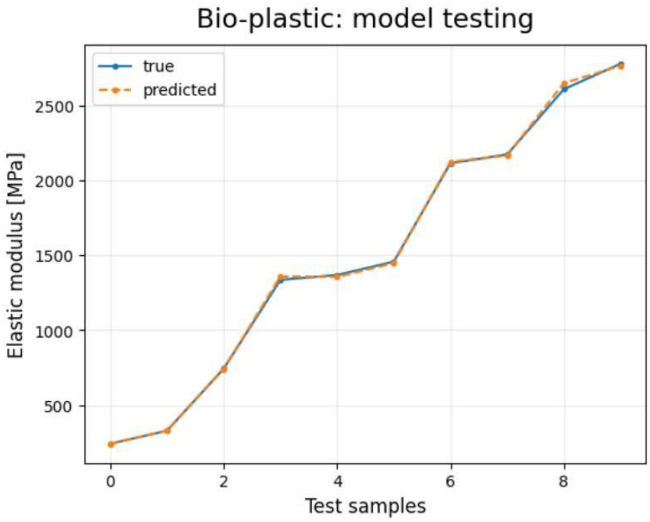


Figure 13. Bio-plastic model test predictions

Model	Validation MAE [MPa]	Validation MSE [MPa^2]	R^2
Model 6	8,3247	263,3448	0.998

Table 12. Bio-plastic model test results

In summary, results are quite good considering the scores obtained during the model testing. In general, the model would be able to deliver a prediction value 8 MPa lower or higher in average than the true value of the elastic module for a bio-plastic material. This value is good considering that the average elastic module of the material is 913,085 MPa. As final remark, the R2 value is almost 1, which can support accepting this model as a perfectly valid model to be used to predict the elastic module of bio-plastic material using as predictor variables the percentage in weight of the additives for the final composition, and distribution of carbon, hydrogen and oxygen of the final composition.

5.3 Wood composites

The third material studied in this report is wood composites, which are targeted for use in manufacturing sliding bearings. These components are widely utilized across industrial, automotive, and heavy machinery sectors. Specifically, in GREEN LOOP, the objective is to replace traditional polymer-based sliding bearings with wood composite-based alternatives within the appliance and tools industries. This shift in the manufacturing paradigm will substantially contribute in the reduction of resources consumption and avoiding the utilisation of environmentally hazardous chemicals and metals.

The advantages of replacing polymer- and metal-based materials with wood composites include similar performance characteristics, such as a low friction coefficient, biodegradability, and resistance to corrosion. Additionally, wood composites can further enhance their performance through optimized manufacturing techniques, such as microwave curing systems, which accelerate production, improve process control, and significantly boost energy efficiency. Collectively, these advancements lead to an overall operational and sustainability improvement, reflected in the enhanced mechanical properties of the final material, including increased durability, strength, and potential for customization to specific application needs.

5.2.1 Wood composites: exploratory data analysis

This section outlines the exploratory data analysis conducted for the wood composites material. First, an overview of the data format used for building the machine learning models of wood composites is presented.

The data available for building the wood composites use case was derived from two types of tests: tensile tests and compression tests. In both types of tests, different material samples with specific dimensions were used to assess the variation in elongation of the test samples under different tensile or compressive forces.

Both tensile and compression tests were conducted on three different materials, namely V236, V241, and V270. Four tests were performed for each type of test and for each material, resulting in a total of 24 different laboratory experiments, which produced 24 distinct datasets. These experiments allow to measure how a wood composite material with specific dimensions changes in length when subjected to tension or compression forces. Additionally, this approach makes it possible to distinguish between linear and non-linear relationships between stress and strain and to evaluate the consistency of the experiments given the different tests performed for each material and each type of test.

The materials used in the tests are identified by their codes. Visually, they appear quite similar, although there are slight differences in colour and shape. For each material, the test samples have consistent dimensions, but these dimensions vary between the different materials. The following tables contain images of the test specimens used, showcasing the test samples employed in both compression and tensile tests for all materials.

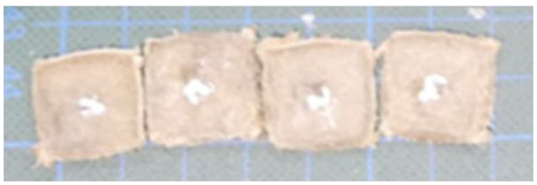
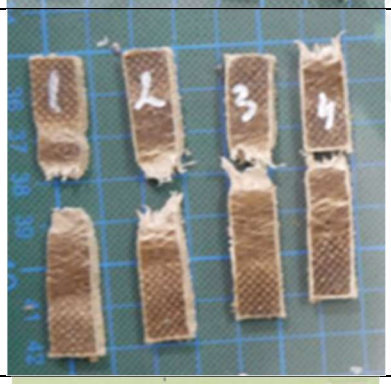
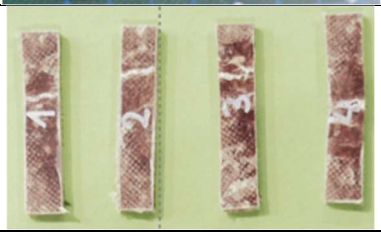
Identifier	Compression test	Tensile test
V236		
V241		
V270		

Table 13. Wood composites, specimens used to perform laboratory tests

The data available for wood composites models contains four variables. The following table summarises the description of each variable contained:

#	Feature	Description	Unit	Feature type
1	Deformation	Percentage of elongation or deformation experienced by the material subjected to tensile or compressive stresses	%	Numerical
2	Tension	Measure of the stress applied to the material in the test, indicating the force per unit area supported by the material	MPa	Numerical
3	Material	Identifier of the material on which the test was performed		Categorical
4	Test type	If the test performed is a tensile or compression test		Categorical
5	Trial	Test identifier for each material and for each type of test		Categorical

Table 14. Wood composites variables

As observed, three out of the five variables in the dataset are related to the test conditions rather than the test results. The variables that describe the evolution of the tests are deformation and tension. In this case, deformation is the variable to be predicted, while tension is the input variable that causes changes in the predicted variable. Thus, the problem is framed as predicting how much wood composite materials deform based on the tension applied to them.

The remaining three variables have different implications for framing the problem. Material, which indicates the type of material, is a fundamental variable as it represents the material to which a specific tension-deformation curve belongs. It is assumed that each material will exhibit different behaviours due to their compositional particularities, so this variable should be considered in later stages of the project. Test type provides essential information about how the test was performed. Essentially, it is a binary variable indicating whether the specimen was subjected to compression or tension. Lastly, trial serves as an identifier that helps distinguish between the different tests performed on each material and test type.

This last variable, trial, does not impact the predicted variable, as it is expected that the material will exhibit similar behaviours in each trial. Additionally, the model is expected to generalize well, meaning no significant variations in behaviour are anticipated between tests. In this way, trial will only be used to split dataset rather than modelling purposes, so after data splitting, trial variable can be discarded.

Given this information, two different modelling approaches can be considered based on the number of input variables included in the model.

The first approach involves creating a single model capable of handling the variables tension, deformation, material, and test type. This model would be more complex, requiring a larger dataset to achieve good performance and to capture the interactions between the different variables. The second approach divides the general problem into smaller, more manageable subproblems, each with a specific model designed to perform well in its particular domain. One way to configure this approach is by creating a separate model for each combination of material and test type, with a high-level selector to recognize the input conditions and route them to the appropriate model. This method significantly reduces the number of variables used for inference, as each model only needs to account for two values: tension and deformation.

To implement this second approach, the original dataset is divided into several subsets, each corresponding to a specific subproblem. In total, 24 distinct datasets are created (based on 2 test types, 3 materials, and 4 trials). These new datasets focus exclusively on the tension and deformation variables, simplifying the model-building process. By narrowing the scope in this way, each model can be optimized for the behaviour of a specific material under a specific type of test, allowing for better accuracy and generalization within those constraints.

As a result, the overall system becomes more modular and flexible, leveraging specialized models for specific conditions rather than relying on a single, more generalized approach. There are several advantages linked to this problem including faster training times, better performance with low-size datasets, and improved interpretability derived from a narrower range of input conditions. The main drawback considered for this approach lies on selecting the model for inference, but considering the simple and straightforward model specializations, a conditional model selector could achieve the expected behaviour by just providing the test type and the material.

Concerning the number of samples available for each material and type of test, the following histograms present the data categorized by test type. As observed, the number of samples for the compression tests is

slightly higher than that for the tensile tests. One reason for this discrepancy is that the tensile test data includes measurements of the behaviour of the material between the maximum tension registered and the point of specimen breakage. In contrast, such intermediate measurements are not taken in compression tests. To eliminate ambiguity in model behaviour, these additional variables from the tensile tests were removed.

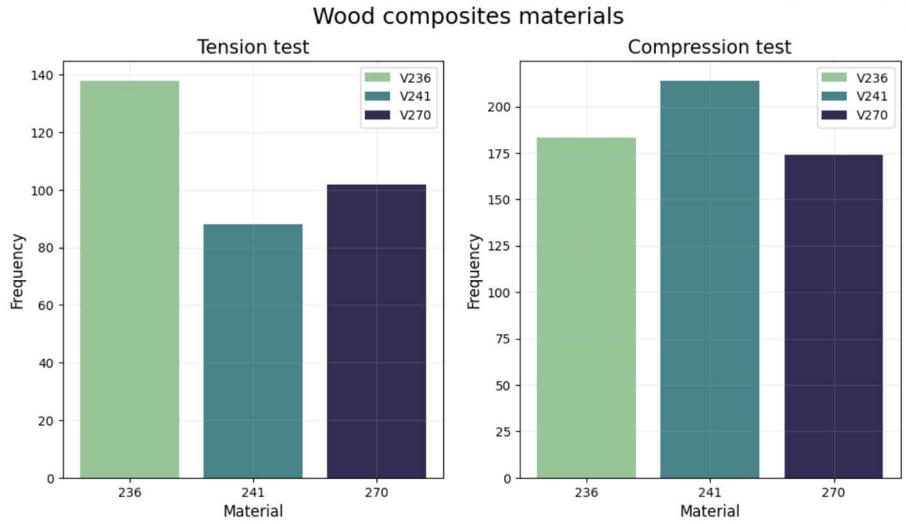


Figure 14. Wood composites: number of samples for each material and type of test

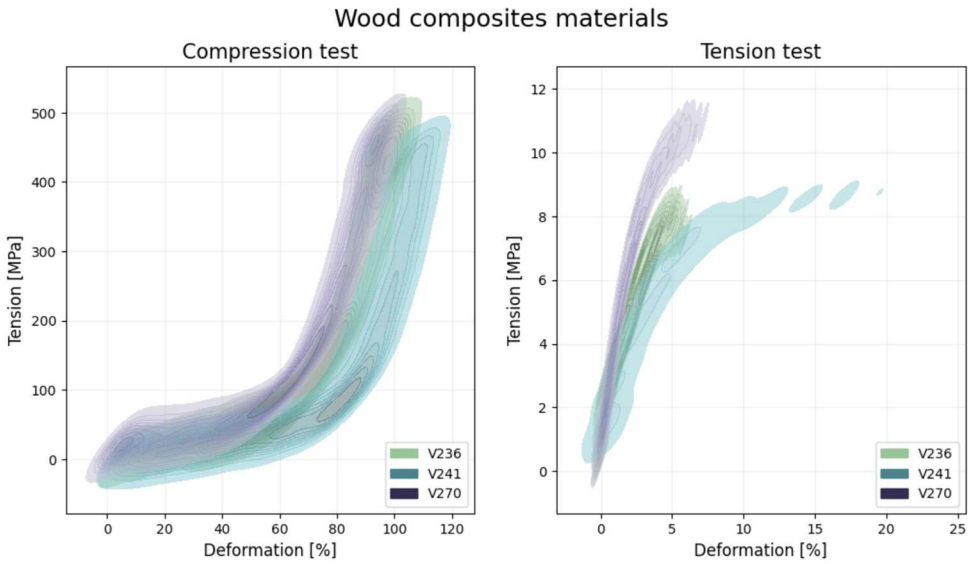


Figure 15. Overall material behaviour for wood composites

The figure above presents a graphical representation of the data discussed and illustrates the behaviour that the models should replicate. It is possible to distinguish between the test types (compression or tension), as each plot corresponds to a different test, and between the materials, which are differentiated by colours.

Compression tests are performed under higher tension conditions, with maximum tension values ranging from 467 MPa to 490 MPa. These values are significantly higher compared to those observed in tension tests, where the maximum tension ranges from 8.3 MPa to 10.92 MPa. The deformations produced vary according

to the material being tested. It is clear that each material exhibits different stiffness levels in both the compression and tension tests. The material with the greatest deformation is V241, followed by V236, and lastly V270. Additionally, the deformations observed in the tension tests are much smaller compared to those in the compression tests. In tension tests, specimens can reduce their size by half in the dimension where force is applied, whereas in compression tests, the maximum dimensional variation observed is 22.2% for V241, 7.19% for V270, and 5.93% for V236.

Considering the information presented in the graph, the implications of performing this kind of study are not limited to determining the elasticity modulus of each material but also extend to understanding the performance of the material under load. This is important because elasticity and resistance to tensile forces are not necessarily correlated: a material may exhibit high elasticity, allowing it to deform significantly under stress, yet still possess low tensile strength, making it prone to failure under relatively low loads. Conversely, a material with lower elasticity may resist higher tensile forces without significant deformation. Therefore, both properties must be considered when evaluating the suitability of a material for specific applications. Finally, main variable descriptors are presented for the wood composites data. Results are provided considering the mentioned problem approach, i.e, dividing the dataset by material and type of test.

	Test	Compression			Tension		
	Material	V236	V241	V270	V236	V241	V270
Deformation [%]	mean	67.148	68.825	57.041	2.240	6.186	2.371
	std	27.283	32.315	29.550	1.667	5.848	1.891
	min	4.014	2.064	0.634	0.007	0.242	0.012
	25%	47.178	40.766	32.051	0.715	1.502	0.790
	50%	74.096	78.236	63.342	2.116	4.217	1.953
	75%	91.499	97.193	83.537	3.554	9.282	3.673
	max	102.526	112.084	95.996	5.937	22.200	7.195
Tension [MPa]	mean	182.522	138.849	171.690	4.790	5.514	6.240
	std	157.777	142.734	150.643	2.262	2.634	3.201
	min	1.487	0.000	14.380	0.350	0.684	0.459
	25%	44.875	22.011	39.946	2.862	3.300	3.459
	50%	125.454	72.970	107.057	5.014	6.040	6.566
	75%	327.272	242.818	290.912	6.804	7.948	9.136
	max	483.966	467.391	490.545	8.307	8.995	10.929

Table 15. Wood composites dataset description

5.2.2 Preprocessing bio-rubber data

The previous section outlined the overall characteristics of the datasets used for the wood composites material. This section details the processes and transformations undertaken to obtain quality data that will be utilized for building, training, and testing the ML models.

During the data cleaning stage, two fundamental tasks were performed. The first task involved removing erroneous entries from the raw data, which were identified as problematic records from the data collection process. Specifically, any tension or deformation values below zero were deemed invalid and subsequently eliminated from the processed dataset. The second cleaning step focused specifically on the tension test data. During data collection, the entire deformation range was recorded, including values beyond the maximum tension applied. After reaching this critical tension threshold, while deformation values continued to increase, tension values began to decrease until material breakage occurred. This phenomenon resulted in multiple deformation values corresponding to a single tension value, potentially introducing ambiguity for the models. Moreover, since there is only one predictor variable for modelling, it is impossible to determine the deformation a material experiences based solely on the applied tension. To address this issue, the approach involved limiting the dataset to only include values up to the maximum registered tension, effectively removing any data collected beyond this point. This strategy ensured consistency in predictions and facilitated the development of ML models that operate effectively within a defined range of values. From a practical perspective, being able to predict material behaviour within this range is particularly valuable, as it reflects the conditions under which the material is expected to perform. Beyond the maximum tensile value, the material is likely to be damaged and its service life is likely to be over.

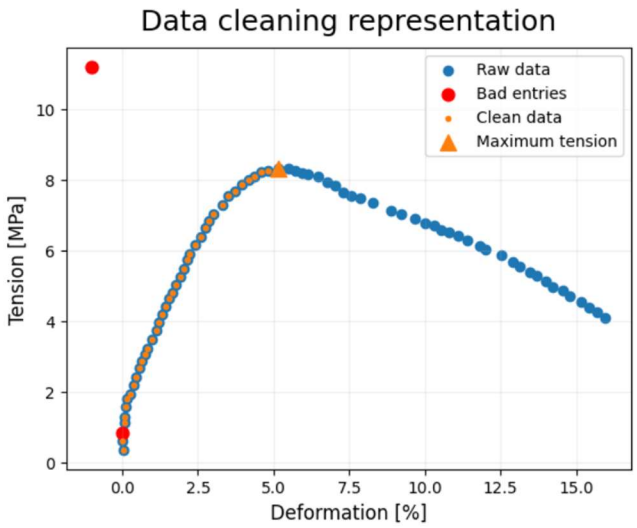


Figure 16. Data cleaning representation for wood composites

The above figure shows the visual representation of the data cleaning process described and performed. The raw data entries are marked in blue. In this case, data entries whose deformation value is negative are marked in red and they have been eliminated. Similarly, data entries whose deformation is beyond the maximum tensile effort suppose a duplication in the tension mapping, and they are excluded for the modelling process. Only points marked with orange pass to further modelling stages.

Once data series are clean, a feature selection process is carried out. As mentioned before, variables present in raw data will occupy different roles during the modelling process. Specifically, deformation and tension

are the numerical variables that ML models will manage. Other variables such as material and test type will be used to define the model domain and target a specific model when inferences are performed rather than being directly used to make predictions. And finally, trial will solely be used to perform the data splitting. Considering that 4 trials are available, data from one of them will be selected as test data and the remaining three trials will be used for training purposes. Once data splitting is done, variable trial will no longer be used in the project. The following table summarises the feature selection process performed:

Variable	Type
Tension	Input / feature
Deformation	Output / target
Material	Model domain
Type	Model domain
Trial	Data splitting

Table 16. Feature selection and role for wood composites

According to these specifications, the data splitting process uses 1 out of 4 trials to divide the data into test and training sets. This results in 75% of the data being used for training and 25% for testing. The data processing step concludes with data standardization, which has been applied only to the input variables. There are two main reasons for this: First, when input variables have different scales, standardization ensures that all variables are on the same magnitude scale. This helps the algorithms balance the contribution of each variable in the prediction process more effectively. Second, the modelling approach used here does not require standardization of the target variable. The algorithm selected (KNN), is an instance-based algorithm and does not perform explicit training like other typical machine learning algorithms. Therefore, the target variable remains unstandardized. After performing data standardization, the data is ready for training the ML models. The results of the standardization are shown below through probability density functions of the input and output data, for both the tension and compression tests, as well as for each of the materials considered in the study. As figures show, each material and each test has its own particularities that are evident through the differences and variations in the shape of the probability density functions.

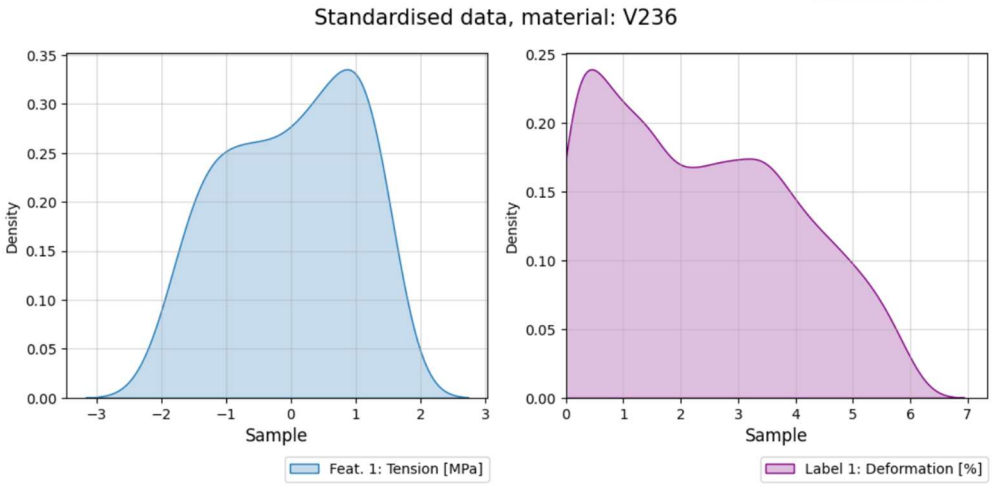


Figure 17. Standardised data for V236 and tension test

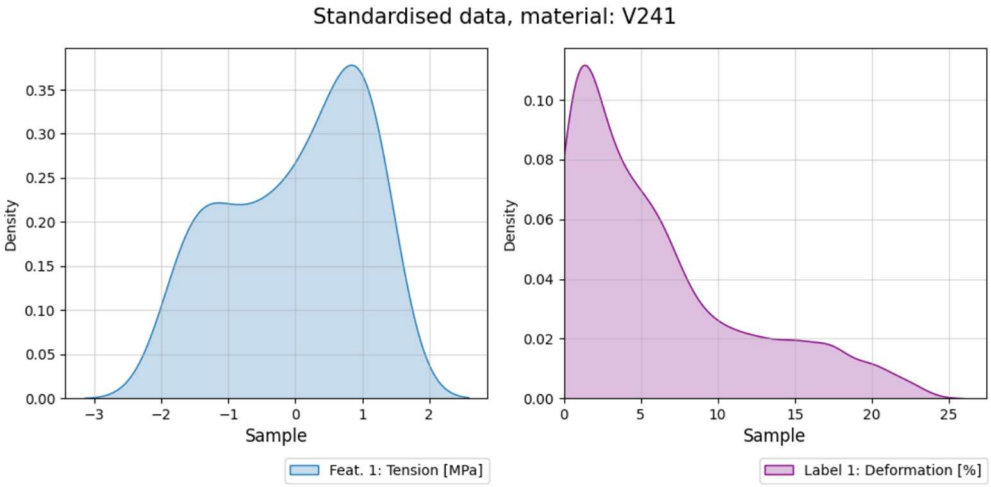


Figure 18. Standardised data for V241 and tension test

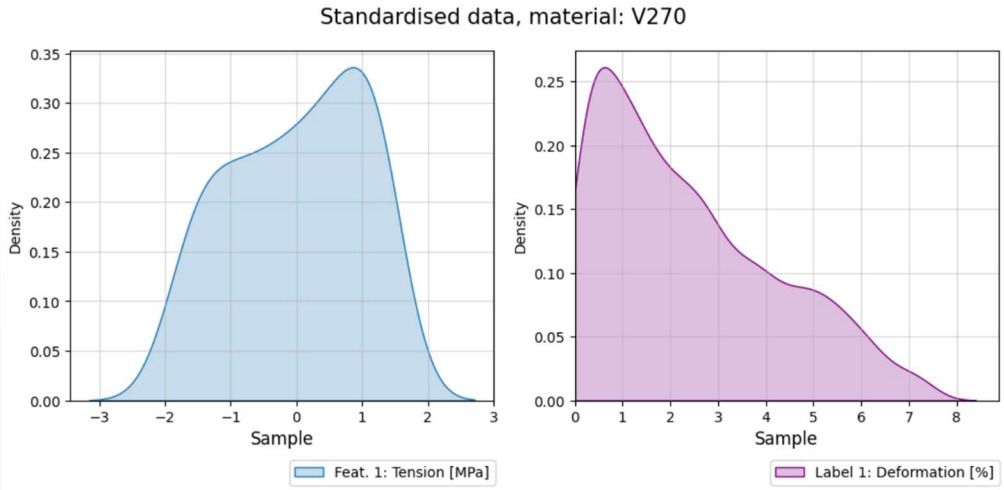


Figure 19. Standardised data for V270 and tension test

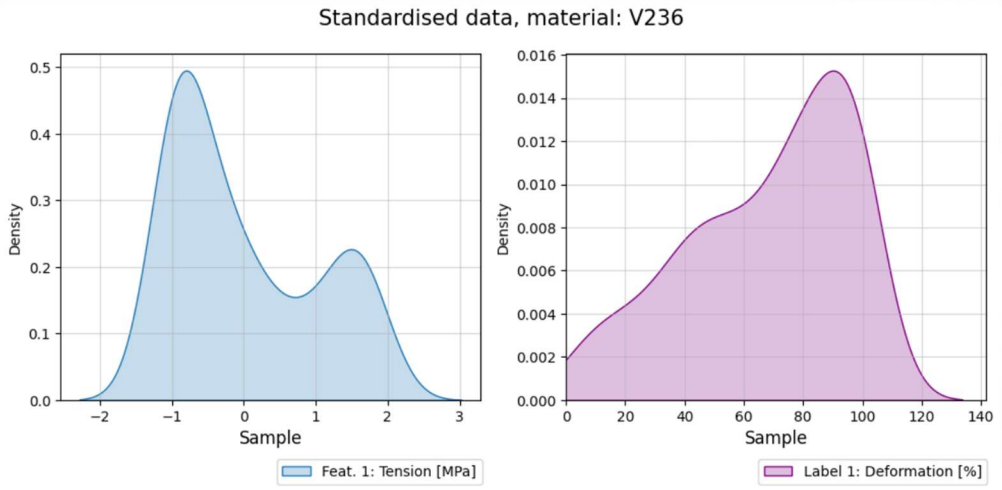


Figure 20. Standardised data for V236 and compression test

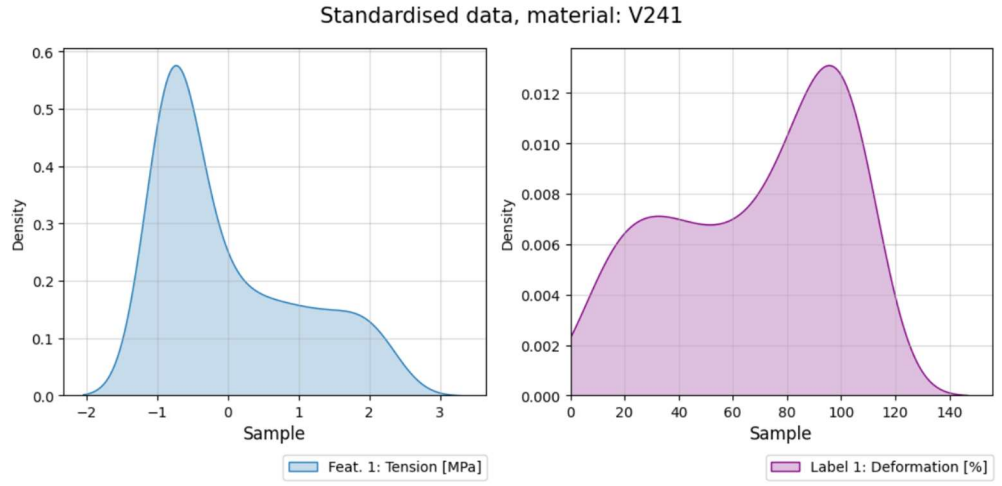


Figure 21. Standardised data for V241 and compression test

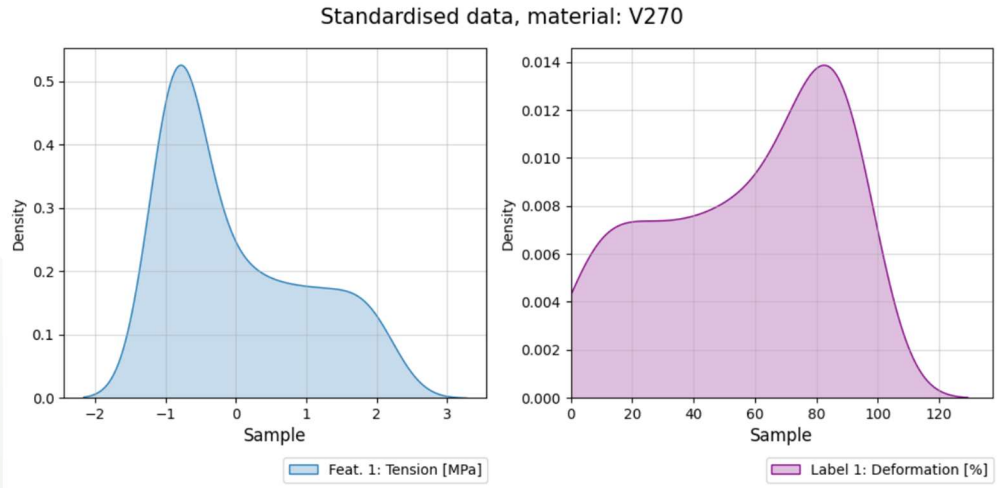


Figure 22. Standardised data for V270 and compression test

5.2.3 Building wood composites model

When approaching the modelling stage, several considerations were taken into account. Given the available data format and the variables involved, the decision to create models specialized in specific domains was a key factor in the algorithm development process.

The models built are tailored to specific domains, meaning they specialize in a particular material and test type—either a tension or compression test. As a result, six distinct ML models will be created to address all possible combinations. Three materials and two types of tests yield six specialized models. To complement this approach, a top-level model selector feature is implemented. This selector uses conditional logic to load the appropriate model for making predictions based on the input data.

For the individual models, various algorithms were evaluated, including polynomial regressors, neural networks, and KNN regressors. After balancing simplicity and performance, the KNN algorithm was chosen. KNN is generally simpler compared to more complex algorithms like artificial neural networks, yet it provides solid performance for this type of problem. KNN requires tuning several hyperparameters before the training stage. In this case, three main hyperparameters were optimized:

- *Number of neighbours*: it defines how many neighbours are considered when making predictions. For a new prediction, the KNN algorithm looks at the n nearest points in the training data. It is crucial to find the optimal value for this parameter, as too few neighbours make the model sensitive to noise and prone to overfitting, while too many neighbours may result in underfitting.
- *Weights*: it determines how much influence each neighbour has on the prediction. Two options were evaluated: uniform, where all neighbours are weighted equally, and distance, where closer neighbours have more influence on the prediction than those farther away.
- *Algorithm*: This specifies the method used to compute the nearest neighbours. The options considered were Ball Tree, KD-Tree, and brute-force search, each with different advantages depending on the characteristics of the dataset (size and dimensionality). The choice of algorithm impacts computational performance, both in terms of model training and inference speed.

In order to assess the most suitable hyperparameter configuration, a regression metric was employed to measure the model performance for the different model configurations. In this case, R^2 was used as the most suitable metric due to its ability to explain the proportion of the variance in the target variable that is predictable from the input features.

Considering this set of hyperparameters and performance evaluation method, a specific KNN regressor was configured for each particular prediction domain. For each of them a specific optimisation was carried out, yielding the most suitable model configuration given the features of the available data. The way in which parameter optimisation was carried out was through grid search. To perform such kind of optimisation, all the model configuration possibilities are tried, and performance results stored. Additionally, the optimization process was carried out using cross-validation. This splitting strategy ensures that model evaluation is done in a practical and robust manner, as models are tested on data they have never seen before. Specifically, the training dataset was divided into k=5 equal parts, where the model is trained on k-1 parts and its performance is assessed on the remaining part. The process is repeated k times, and in each time, a different part of the original dataset is used as test set. In this way, hyperparameter combinations are adjusted in the most general way, meaning that decision does not depend on a specific data splitting that can inherit specific data skews, but in the average of the k iterative data splits performed.

The following figure displays the results of the optimisation activities, and it shows the evolution of the performance metric as the number of neighbours increases.

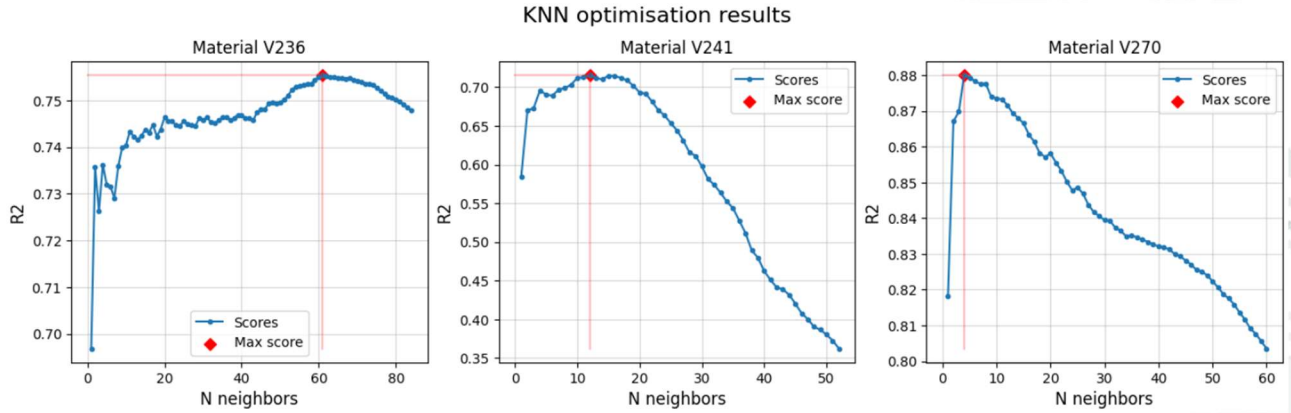


Figure 23. KNN optimisation for compression

Material	Number of neighbours	Weights	Algorithm
V236	61	Uniform	Ball Tree
V241	12	Uniform	Ball Tree
V270	4	Uniform	Ball Tree

Table 17. Best KNN hyperparameter configuration for tension test

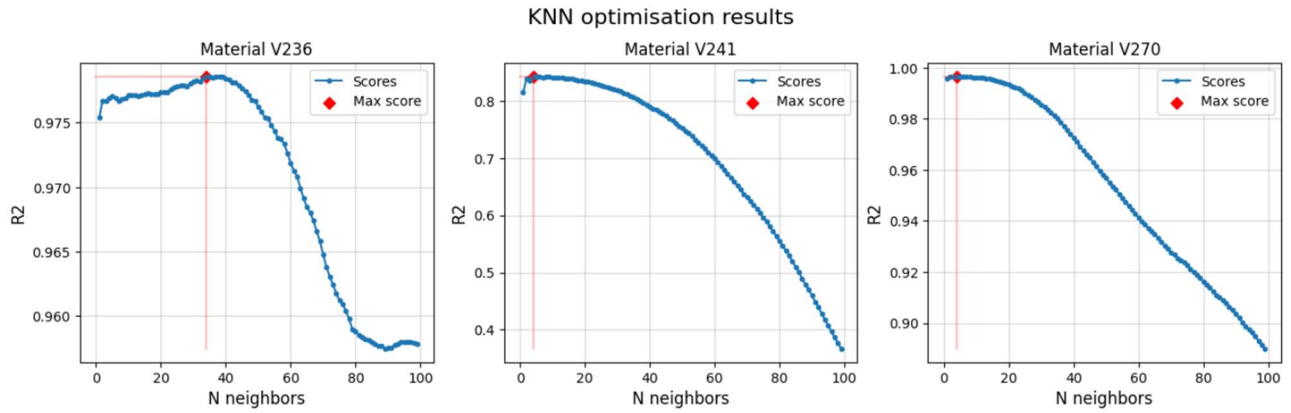


Figure 24. KNN optimisation for tension

Material	Number of neighbours	Weights	Algorithm
V236	34	Uniform	Ball Tree
V241	4	Uniform	Ball Tree
V270	4	Uniform	Ball Tree

Table 18. Best KNN hyperparameter configuration for tension test

As can be observed, the results show high variability in the number of neighbours, ranging from 4 to 61, but a consistent choice of weights and algorithm, with uniform weighting and the Ball Tree algorithm being selected across models. These conclusions are influenced by the characteristics of the data, which are themselves derived from the properties of the materials. More uniform materials generate more consistent datasets, leading to higher-quality data. In contrast, less uniform materials tend to behave more variably, producing noisier data that complicates the modelling and prediction tasks.

5.2.4 Training wood composites model

With the results displayed for hyperparameter optimization, the different models were created and trained. In this case, since KNN is an instance-based algorithm, no model parameter fitting was required. Instead, the model is instantiated, and the data (both features and labels) are loaded, allowing the algorithm to internally store the training examples. When a prediction is made, the algorithm will compare the new input with the stored data, identifying the k nearest neighbours based on the defined distance metric. The prediction is then calculated by averaging the target values of these neighbours, weighted according to the uniform scheme.

5.2.5 Testing bio-rubber model

Finally, the wood composites models were tested using the data from the trial designated for testing. Although this dataset follows similar patterns to the training data, it has never been seen by the model, allowing for a thorough assessment of both model performance and generalization capabilities.

As with earlier evaluations, three different metrics were employed to measure model performance across different dimensions: MAE, MSE, and R^2. This approach ensures a comprehensive evaluation of the regression models, providing insights into their ability to minimize both small and large errors (through MAE and MSE, respectively) and their overall explanatory power (via R^2).

Purpose	Material	MAE [%]	MSE [%]	R^2
Tension	V236	0,4077	0,2335	0,9132
Tension	V241	0,8875	1,4559	0,9395
Tension	V270	0,1866	0,1165	0,9748
Compression	V236	1,9227	5,861	0,9912
Compression	V241	2,6015	8,7055	0,9918
Compression	V270	0,6669	0,8473	0,999

Table 19. Wood composites model test results

As the table shows, the models demonstrate high performance across all evaluated metrics. Specifically, for the R^2 metric, test results yield values very close to 1, indicating that the models are able to explain almost all of the variance in the target variable. An R^2 value near 1 suggests a strong correlation between the predicted and actual values, meaning the model captures nearly all relevant patterns in the data with minimal unexplained variance. Notably, the results are particularly strong in the case of compression tests. While the MAE and MSE values are slightly higher than those for the tension tests, this can be attributed to the broader range of deformation values observed in compression, where the model must predict deformations approaching 100% of the initial length of the material. Despite this larger prediction range, the compression models still exhibit exceptional accuracy, as reflected by R^2 values exceeding 0.99 across all materials. This

high R^2 indicates that the models not only perform well within the broader prediction range but also maintain an extremely low margin of error, ensuring reliable and precise predictions.

After the internal testing, all the code, including the KNN models and model selector, was embedded into an executable program along with a recipe specifying the steps required to perform predictions. This program was sent to Fraunhofer for external validation. As of now, the model has not yet been validated. Once feedback from Fraunhofer has been received, wood composited models will be entirely validated.

6 Prediction of temperature profile through convolutional neural network

Due to the increasing integration of microwave heating into industrial processes, there is a growing demand for precise temperature control and prediction within materials. The relationship between microwave power and material temperature involves a complex interplay of electromagnetic flux, material properties, and geometric configurations. Traditional methods to compute this relationship rely on numerical simulators, such as finite element analysis and computational electromagnetic modelling, which typically yield good results in terms of performance and accuracy. Consequently, researchers and engineers use these tools to optimize process parameters and study how to achieve uniform heating while minimizing undesired effects like hot spots and thermal gradients [18], [19]. However, these tools come with significant computational costs, making them time-consuming and resource intensive.

Several studies support the use of ML techniques for predicting temperature in microwave systems with high accuracy [20], [21], [22]. This study builds on that line of research by applying convolutional neural networks to predict temperature in MW heating systems. The work involves applying CNNs to datasets generated from microwave heat simulations [23] to analyse the behaviour of various industrial materials, including susceptors, semi-transparent, and transparent microwave materials. Specifically, the models are used to predict temperature profiles and heating times under different power conditions measured through minimum, average and maximum temperatures. Trials were made for 9 materials under identical conditions. These conditions consisted in microwave sourcing from 100 W to 500W, identical geometry and initial temperature of 25ºC.

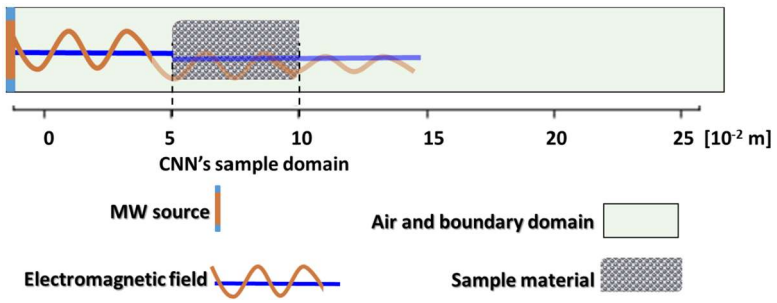


Figure 25. Graphical description of microwave system trials

6.1 Temperature profile from microwave heating: EDA

The data utilised in this case has been generated through numerical simulators. Unlike other use cases in which data records meant a static value or the result of a specific test, in this case, data is in time-sequence format, meaning that variables are associated to a specific time stamp. This means that they follow a hierarchy in the tabular format, and that the values of the predicted variable not only depend on the input variables but also on the previous values of the predicted variable.

#	Feature	Description	Unit	Feature type
1	Sample	Sample identifier		Categorical
2	Power	Microwave power applied to the test sample	W	Numerical
3	Time	Time frame in which the record was registered	s	Numerical
4	Min temperature	Minimum temperature recorded in the material during the test	ºC	Numerical
5	Average temperature	Average temperature recorded in the material during the test	ºC	Numerical
6	Max temperature	Maximum temperature recorded in the material during the test	ºC	Numerical

Table 20. Microwave temperature profile variables

Materials utilised in this study can be classified in three major categories. For each category, different materials were tested. Details about the categories and materials utilized are summarised in the following table:

#	Material	Category	Description
1	ALN Powder compact	Susceptors	Susceptor material characterised by its high thermal conductivity and electrical insulation properties
2	CuO Powder compact	Susceptors	Susceptor material utilized in microwave-assisted processes for absorbing microwave energy
3	SiC	Susceptors	Susceptor material used in high-temperature applications due to its high thermal conductivity, hardness and ability o absorb electromagnetic radiation
4	Soda Lime-Glass	Semi-transparent	Common glass type used in varied applications (windows, bottles, …) with good chemical durability and low thermal resistance
5	Alumina Silicate	Semi-transparent	Ceramic material made from aluminium oxide and silicon dioxide characterised by its heat resistance and semi-transparent properties
6	Borosilicate-Glass	Semi-transparent	Material characterised by its low thermal expansion that endows it with high resistance to thermal shock
7	Alumina cement	Semi-transparent	It is a high-performance cement made from calcium aluminates, often used in refractory applications
8	Boron nitrate	Transparent	Synthetic ceramic material known for its high thermal conductivity, electrical insulation, and lubrication properties
9	Dense mullite	Transparent	Ceramic material formed from aluminium silicate, known for its thermal stability and used in high-temperature environments

Table 21. Microwave temperature materials

As previously mentioned, the dataset used in this study was generated from numerical simulations. An example of one simulation result is presented to illustrate the data used to build the model. As the data shows, power gradually increases from 100 W to 2000 W, while temperatures evolve differently across the material. The minimum temperature corresponds to areas that are difficult to heat due to dimensional constraints or wave orientation issues. These regions take longer to reach higher temperatures, which

explains why the minimum temperature rises only after a delay. In contrast, the maximum temperature occurs in areas where microwave energy is directly focused, causing them to heat up more quickly.

Sample	Power [W]	Time [s]	Minimum temperature [ºC]	Average temperature [ºC]	Maximum temperature [ºC]
1	100	0,028	25	25,016	25,048
2	100	0,055	25	25,023	25,087
3	100	0,083	25	25,048	25,120
4	100	0,111	25	25,064	25,154
…	…	…	…	…	…
215588	2000	3,9986	204,860	347,9283	552,9065
215589	2000	3,9993	204,9478	348,0554	553,0815
215590	2000	3,9999	205,0356	348,1824	553,2565

Table 22. Microwave heating dataset example

6.2 Preprocessing

The preprocessing steps applied to the dataset focused on formatting it into appropriate time series with a specific window length and set of variables. The window length, which defines how much past data the model processes to make predictions, plays a crucial role in the performance of the model. A longer window provides more past information but could lead to overfitting or increased computational costs, while a shorter window might not capture enough context. Initially, the data was presented as a single table, so it had to be split and windowed to convert it into the proper time-series format.

The transformations for inputs and outputs followed distinct processes. For the inputs, the data was first windowed based on the window length, and relevant variables were selected. These variables were then combined to form a two-dimensional input tensor representing time and microwave power. Afterward, the inputs were standardized to ensure compatibility across different ranges, as the variables had different units and magnitudes.

For the outputs, which were temperature readings, a vector was created that captured the minimum, average, and maximum temperatures. The label for each input sequence was the temperature values at the last time step in the window. In other words, if the input window spanned the first 30 values in a sequence, the label would be the 30th value of the minimum, average, and maximum temperatures. This approach ensures that the model learns to predict the future temperature states based on a sequence of historical data.

6.3 Building model

The model built for this application has been developed following some guidelines imposed by the type and format of the data managed and the objective of the study itself. Firstly, due to the temporal format of the dataset, it is expected that data is ingested in sequence format rather than vector format. Secondly, the model will be used to get the minimum, average and maximum values of temperatures registered in the material. In this way, the model should be suitable for working with both time-series data as with vectorial data. In this sense, CNNs appears as a powerful AI technique able to capture complex temporal micro-wave induced patterns. Although their most frequent use is in the field of computer vision, their application extends far beyond, and they show proficient capabilities when managing temporal sequences. The reason is that they are able to efficiently manage multidimensional data, so treating time as a spatial dimension, CNNs are able to unravel the sequential dependencies and features within the simulated microwave heating data.

The basic operation in which CNNs base their work is the convolution. This mathematical operation processes data by sliding a smaller array known as kernel or filter over the input tensor:

$$g(x,y) = k \cdot f(x,y) = \sum_{i=-a}^{a} \sum_{j=-b}^{b} f(x-i, y-j) \cdot k(i,j)$$

Where:

- $g(x,y)$: is the convolution function applied at each x and y position in the input tensor.
- k: is the kernel used in the function.
- $f(x,y)$: input tensor.
- a: number of rows considered in the convolution function.
- b: number of columns considered in the convolution function.

In this neural network architecture, neurons serve as the computational units that process information extracted through convolution. They apply a filter to their receptive fields and produce a raw output value that represents a feature or a part of a potential feature within that region. The receptive fields are specific areas of the input feature map that neurons "see" and process, allowing them to extract local patterns within their respective regions.

After the convolutional operation, activation functions are applied to capture non-linearities present in data and endowing the model with the capacity to reproduce complex patterns. In this way, the activation functions decides whether and to what extent a signal should progress further through the network, effectively determining the activation of the neuron based on the input received.

The results of these sequential mathematical transformations create new feature maps with a higher level of abstraction. These feature maps represent the input data with specific information emphasized by the applied filters.

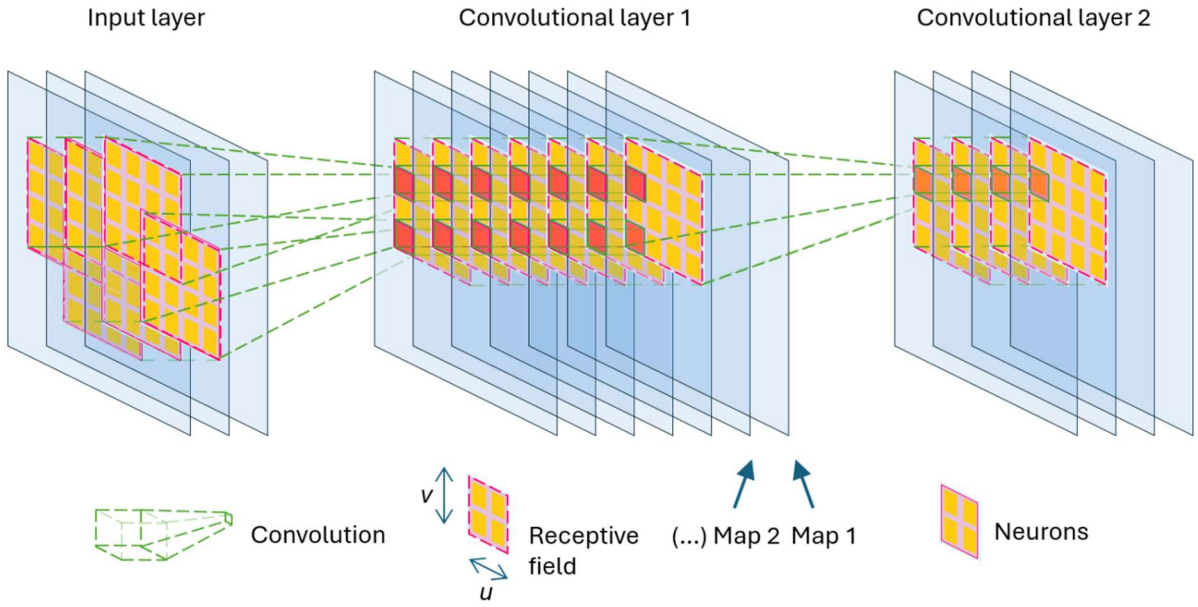


Figure 26. Convolutional layer representation

In the figure, data flows from left to right. Each neuron in a feature map of layer *n* receives information from its receptive field in layer *n-1*. This flow of information is depicted by the dashed green lines, which also represent the convolution operation. The number of feature maps in layer *n* corresponds to the number of kernels, or filters, applied during convolution to the feature maps of layer *n-1*.

Another fundamental component in CNN architectures are pooling layers. They are used to reduce spatial dimension of feature maps generated by convolutional layers. They perform down-sampling operations to reduce the number of parameters and computational complexity while also making the features extracted by the convolutional layers more robust to variations in the position of features in the input. Dimensions of pooling layers are specified through the pool size parameter. In this way, it is possible to define the region over which the pooling operation is applied. The most common pooling operations are max pooling or average pooling, in which the pooling layer extracts the maximum value of the pooling region or the average value of the pooling region respectively.

When building a CNN, pooling layers are typically inserted after each convolutional block. A common approach is to stack multiple convolutional layers and complete the block with a pooling layer. This approach offers several advantages: i) Pooling layers reduce the spatial dimensions of the feature maps, thereby lowering the computational load and reducing the risk of overfitting by focusing on the most critical features; ii) Pooling layers enhance the invariance of the network to small shifts, distortions, or noise in the input data. This contributes to the ability of the model to generalize more effectively, improving its capacity to recognize patterns across diverse datasets.

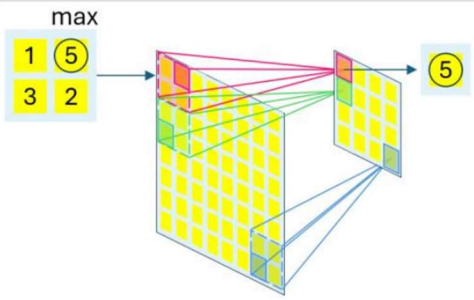


Figure 27. Max pooling layer representation

When building neural networks with convolutional blocks, it is common to include a head block composed of dense or fully connected layers, depending on the application. This architecture leverages the strengths of both types of layers: convolutional layers excel at spatial feature extraction, while dense layers enhance pattern recognition. However, these layers handle data differently: convolutional layers work with multi-dimensional data, while dense layers require one-dimensional tensors. At this stage, the tensor produced by the network is three-dimensional, so a dimensional reduction is necessary. This is achieved using a flattening layer, placed between the convolutional and dense blocks, which converts the 3D tensor into a format suitable for input into the dense layers. The following figure depicts the connection between both blocks and the representation of the flatten layer:

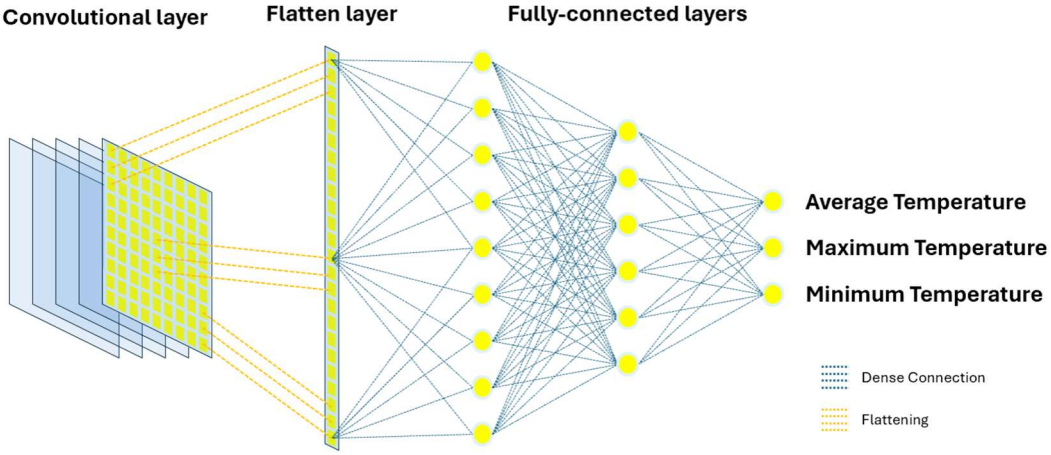


Figure 28. CNN top structure: transition from CNN to head block

In order to optimise the model architecture and configuration, different hyperparameters were tested. The first hyperparameter tested was the activation functions employed. The only constrain here is that the activation function of the output layer must be a function without upper limit. The reason for this is that when predicting temperature values for specific materials according to different heating patterns, the actual temperature values should not be limited to a specific value. After trying different activation functions, best results were obtained for ReLU.

Regarding the model architecture, best performance was obtained for this structure:

- Two convolutional layers with 128 and 64 feature maps. Kernel size 2x2.
- Max pooling layer
- Fully connected head block with 3 layers (50, 16 and 3 neurons each layer).

For avoiding overfitting, L2 regularization was included in the dense layers. This constraint penalized strong differences between the weights contained in a single layer, thus improving the generalization.

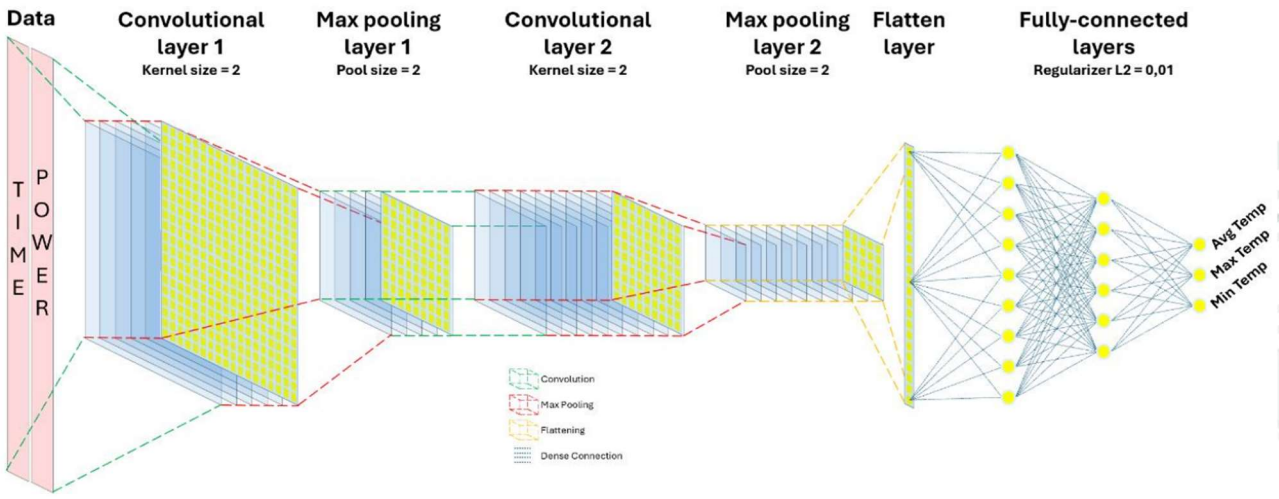


Figure 29. Final model architecture

The figure shows the model created from implementing the above-mentioned architecture. As it was justified, each component in the network fulfils a specific purpose, and its design is optimised to achieve optimal results in forecasting material temperatures using power time-series data.

This network architecture facilitates the extraction and refining of information as data flows through it. In the first block of the layer (until the flatten layer), each layer will apply convolution operation to increase the level of abstraction of data, while the max pooling layer will reduce the spatial dimension of the resulting tensors, thus simplifying the information while retaining the most significant features.

The top block of the network starts with a flatten layer, which receives the resulting tensor output from the bottom block and yields a one-dimensional array of 256 values. Then the information is ready to flow through the dense layers, which, equipped with L2 regularization to prevent overfitting, will finally yield the three temperature values: average, maximum and minimum.

6.4 Model testing

In this section the results obtained during the implementation of the model described in the above section are presented. Results are presented following a hierarchy of type of material (transparent, semi-transparent, and susceptor), material, and output variable (minimum, average and maximum temperatures). For each individual case, model outputs are compared against actual numerical simulation values.

5.2.1 Transparent materials

Boron nitrate

The figure shows the results for minimum, average and maximum temperatures for Boron Nitride. The model demonstrates good alignment with the actual data, indicating a more accurate performance during periods of steeper temperature increase. This is observed mainly when simulating high microwave power. For lower power rates small differences between actual and model prediction are appreciated, although these differences barely exceed 0.2 ºC.

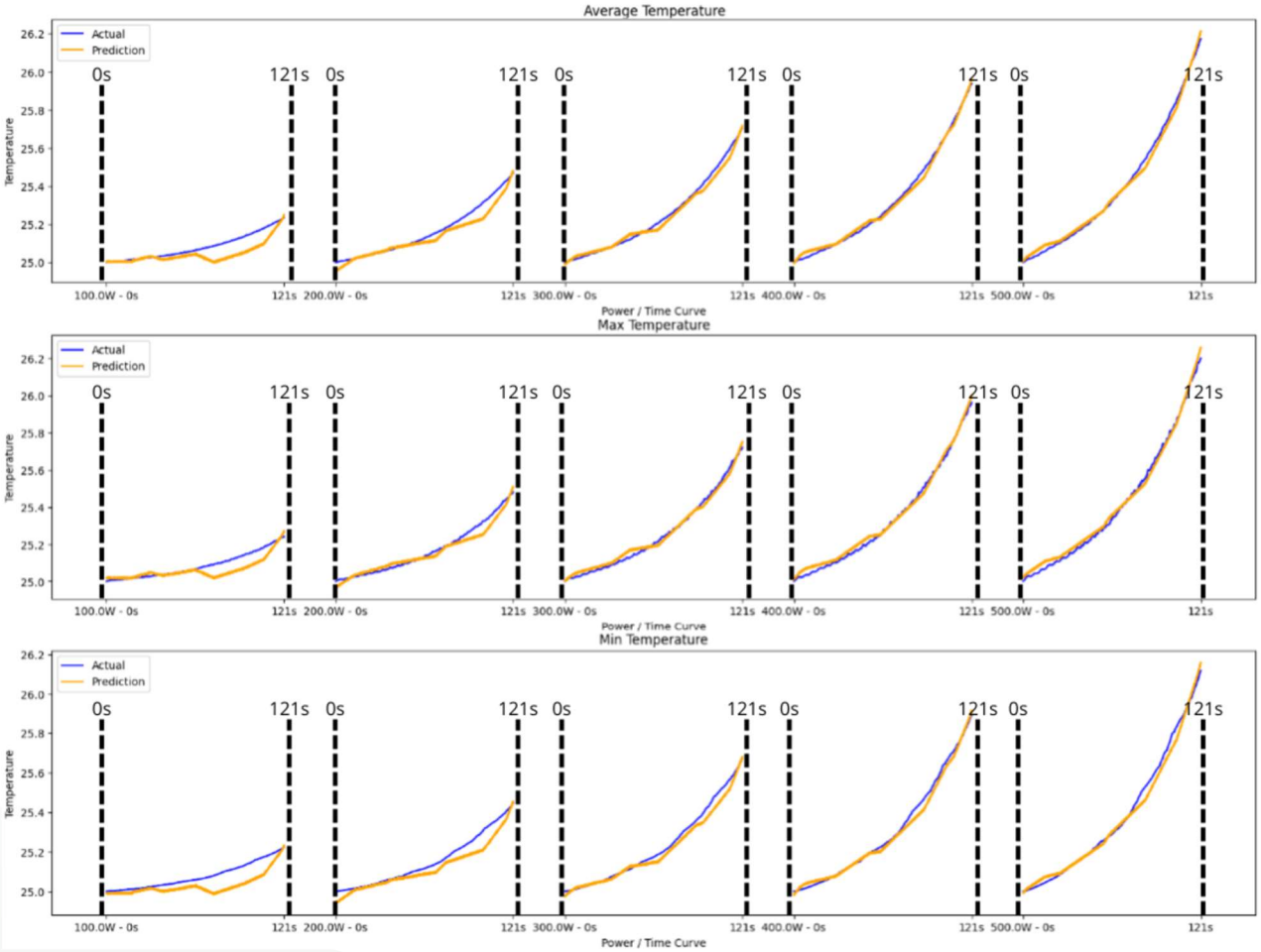


Figure 30. Boron nitrite: graphical testing results

Minimum temperature results: The prediction and actual curves align closely, with the prediction slightly lower in the latter half. This implies that the model might be slightly conservative in estimating temperature increases, possibly due to a high emphasis on the initial thermal characteristics of the material. There is a consistent pattern where the model slightly underestimates the temperature before aligning more closely as time progresses. The initial lag could be due to a delayed thermal response within the model's parameters.

Average temperature results: Both the actual and predicted temperatures show a consistent and gradual increase over time. The model follows the actual temperature trend closely, suggesting it has captured the thermal response of the material to this power setting well. For higher power rates (>300), the prediction closely mirrors the actual upward trajectory of the temperature, with a slightly steeper slope. This may indicate that the model is slightly overestimating the heat capacity or underestimating the heat dissipation rate for the material. However, the convergence of the prediction with the actual values over time suggests the model adjusts well as it receives more data points.

Maximum temperature results: The predictions closely follow the actual temperatures, with a minor overestimation of 0.2 ºC in the mid-section of the curves. This might indicate that the model expects the material to retain heat more than it does. Yet, the model accurately captures the thermal behaviour trend, which is a positive indication of its learning from the response of the material. Min. Temperature Curves The prediction and actual curves align closely, with the prediction slightly lower in the latter half. This implies that the model might be slightly conservative in estimating temperature increases, possibly due to a high emphasis on the initial thermal characteristics of the material. There is a consistent pattern where the model slightly underestimates the temperature before aligning more closely as time progresses. The initial lag could be due to a delayed thermal response within the parameters of the model.

Numerical results obtained from testing the model with boron nitride are presented in the following table:

	Minimum temperature	Average temperature	Maximum temperature
MAE	0,027	0,022	0,024
Max AE	0,290	0,327	0,326
MSE	0,001	0,001	0,001

Table 23. Boron nitrate: numerical test results

The MAE values across the Mean, Max, and Min temperature predictions are remarkably low, with values of 0.022, 0.024, and 0.027, respectively. This indicates a high level of precision in predictions, with an average deviation from actual temperatures being mere hundredths of a degree. Such precision is commendable and indicates the model is well-tuned to predict temperature fluctuations with minimal error. The maximum errors recorded are 0.327, 0.326, and 0.290, while these values are higher than the MAE, it is important to note that they represent the worst-case deviations in the model predictions. Considering the potential variability and complexities involved in predicting temperature changes in different materials, these Max Absolute Errors values suggest that even in the most challenging scenarios, the model maintains a reasonable level of accuracy. The MSE values for all three temperature predictions stand uniformly at 0.001. This metric, emphasizing the square of the errors before averaging, indicates that the model consistently maintains a tight bound on errors, with very few large deviations. A low MSE is particularly impressive as it suggests not only that the average error is low but also that the distribution of errors skews towards smaller, less impactful mistakes.

Dense mullite

The results obtained for dense mullite are depicted in this section. The graphical results display a high fit between temperature curves from both predictions and actual values, reflecting the robustness of the CNN model and its ability to understand the thermal response of the material. Such predictive accuracy is indicative of a well-developed model that captures the nuances of temperature change in response to varied power inputs.

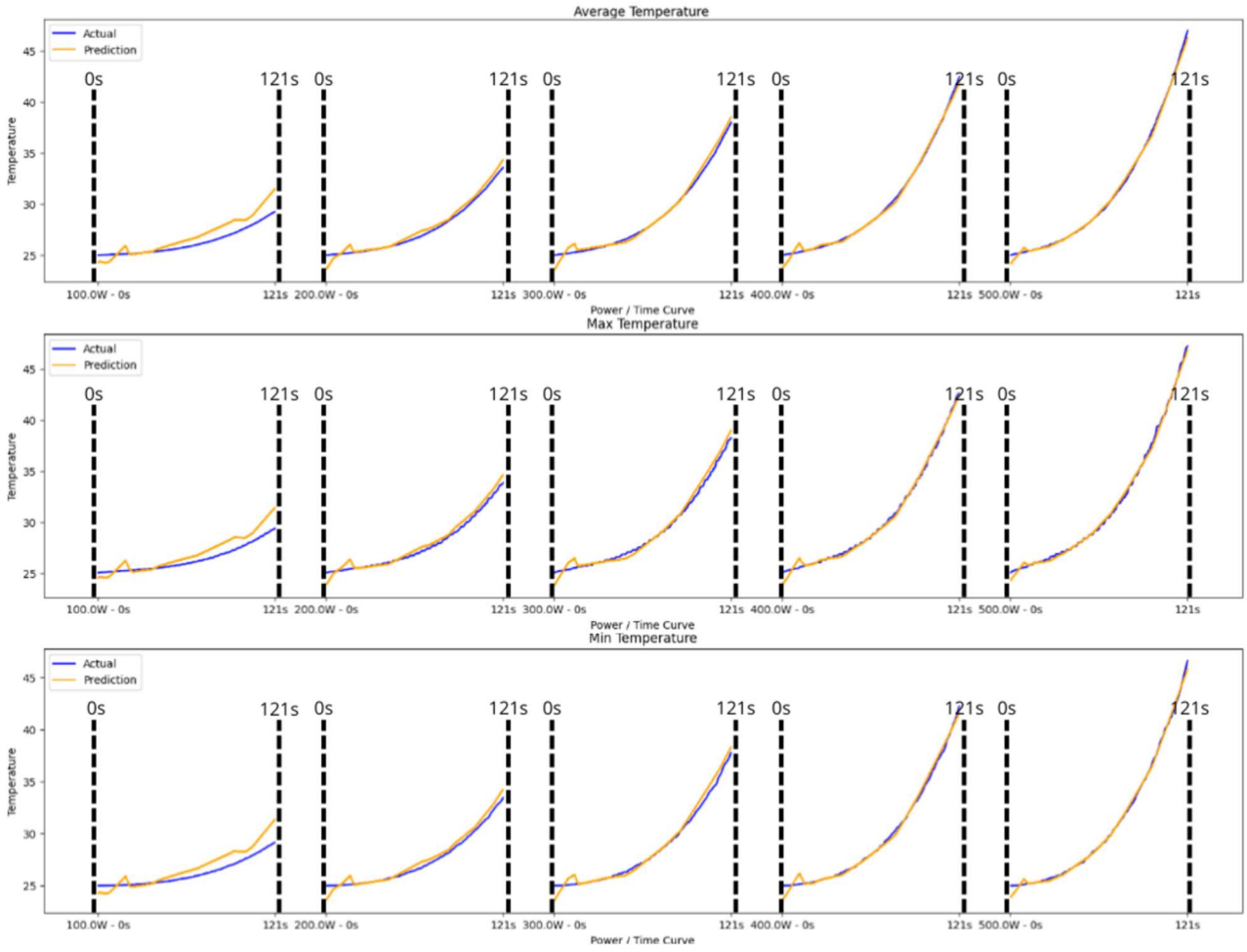


Figure 31. Dense mullite graphical test results

Minimum temperature results: The prediction consistently tracks the actual temperature with high fidelity, indicating the robustness of the model in understanding the response of the material to varying energy inputs. The model seems particularly adept at capturing the rate at which the material cools down or its initial thermal inertia.

Average temperature results: in this case, both the actual and predicted curves begin closely aligned, reflecting the accuracy of the model in the initial stage of the heating. As time progresses, the prediction slightly underestimates the temperature (<1,5 ºC) in low power rates, which could mean the model may be slightly conservative in its heat accumulation estimation for the material at this power level. Across higher power rates, the predictions are very close to the actual temperatures, particularly in the mid-section of the simulation range. The prediction slightly leads the actual temperature, potentially indicating a model

expectation of a quicker thermal response from the material than observed. The close tracking throughout suggests the model has a good grasp of the thermal dynamics at play.

Maximum temperature results: The prediction closely follows the actual curve with impressive accuracy. There is a slight overestimation as the curve progresses (<1.5ºC), hinting that the model might slightly overpredict heat retention in the material at this low power setting. For high-power simulations, the model closely matches the actual temperature trend, with minimal overestimation at higher temperatures. This suggests that the model is well-calibrated to the material's response to increased power levels, though it may slightly overestimate the thermal conductivity or heat capacity of the material.

Numerical results for boron nitrate are presented:

	Minimum temperature	Average temperature	Maximum temperature
MAE	0,623	0,631	0,678
Max AE	13,348	13,837	14,933
MSE	1,225	1,268	1,477

Table 24. Boron nitrate: numerical test results

With MAE values of 0.631 for Mean Temperature, 0.678 for Max Temperature, and 0.623 for Minimum Temperature, the model demonstrates a commendable level of average accuracy across all temperature ranges. These figures suggest that, on average, the temperature predictions deviate from actual measurements by a small margin. Such performance is indicative of a well-constructed model that can reliably capture the general thermal behaviour of the new material.

The Max Absolute Error values, recorded at 13.837 for Mean Temperature, 14.933 for Max Temperature, and 13.348 for Min Temperature, are not uncommon in predictive modelling, especially when dealing with materials that exhibit complex thermal responses.

The MSE values of 1.268 for Mean Temperature, 1.477 for Max Temperature, and 1.225 for Minimum Temperature further reinforce the notion of consistent performance in predictions. Thus, the model demonstrates good predictive performance with all temperatures. The errors are quite low, indicating that the model captures the overall trends and the nuances in the temperature data well. The similar MAE and MSE values across different temperature readings suggest a balanced model that doesn't disproportionately struggle with any aspect of temperature prediction.

5.2.2 Semi-transparent materials

Soda lime-glass

This section presents the results obtained during model testing for soda lime-glass material. In this case, the response of the material to heat seems complex and includes sudden changes driven by non-linearities present in data that challenge the modelling process.

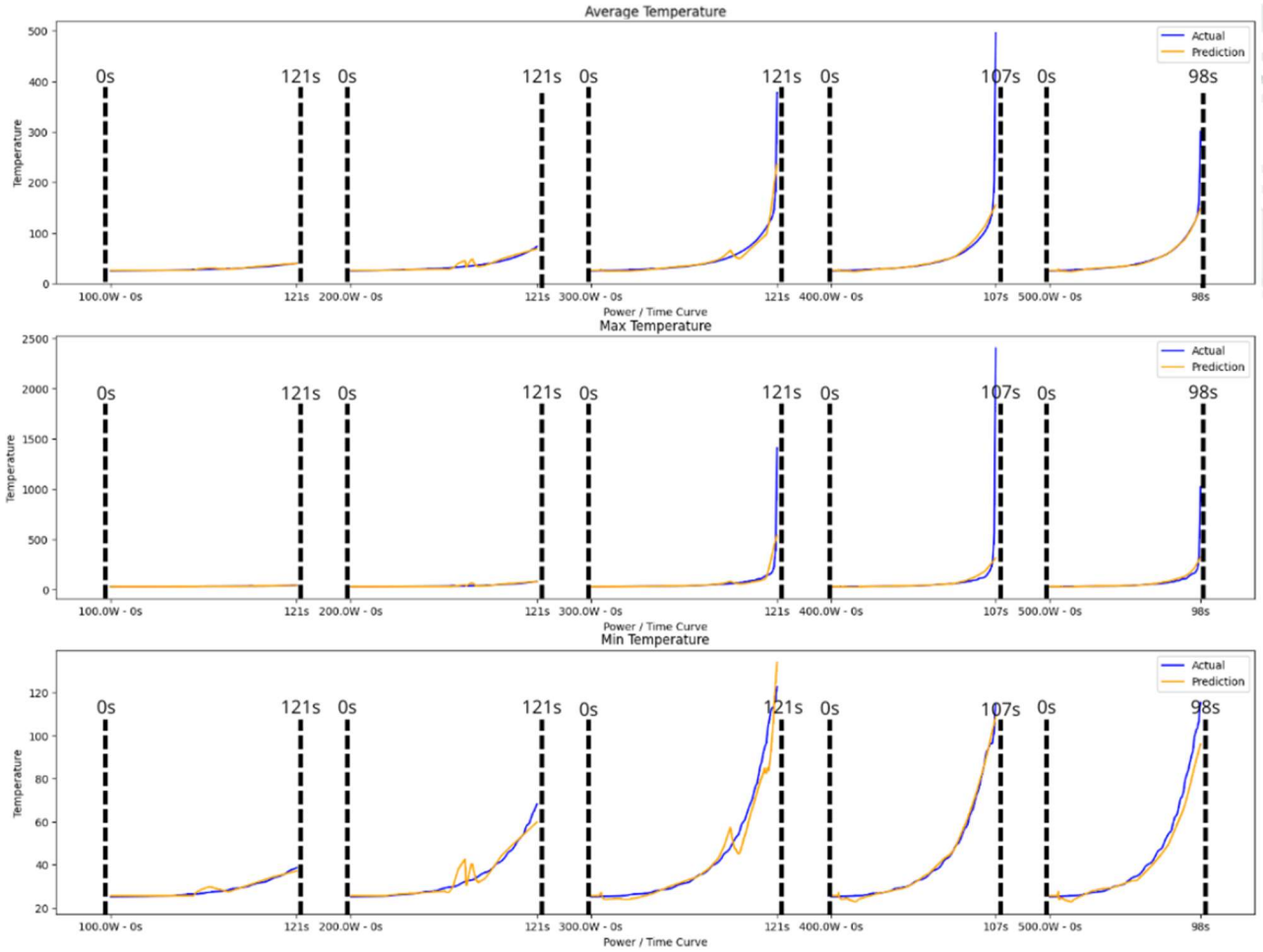


Figure 32. Soda lime-glass graphical test results

Minimum temperature results: the prediction captures the actual trend with high fidelity, indicating the ability of the model to understand the thermal response of the material. Prediction is quite robust at lower to mid-range power settings, but at higher powers the prediction slightly overshoots the actual temperature during the cooling phase, suggesting that the model might overestimate the material's ability to retain heat as the power setting increases.

Average temperature results: the prediction closely mirrors the actual temperature, except for a minor lag at the end. This could mean the model slightly underestimates the rate at which the material reaches its maximum temperature under high power conditions. It can be seen that as the power increases, the model maintains a strong fit with the actual temperature, reflecting its ability to scale its predictions appropriately with the increase in power. There is a slight lag in the prediction at the start, suggesting initial conditions or response time could be further calibrated.

Maximum temperature results: the predictions are well-aligned with the actual temperature, showing only minor deviations. The model seems well-tuned to the thermal properties of the material under these conditions. However, there is a notable discrepancy at the final temperature, where the model underestimates the actual temperature. This suggests that at very high-power levels, the model may not fully account for the heat retention or the rate of temperature increase.

	Minimum temperature	Average temperature	Maximum temperature
MAE	2,160	4,871	23,615
Max AE	33,239	693,082	4034,911
MSE	13,868	666,841	22163,493

Table 25. Soda lime-glass numerical test results

The MAE values of 4.871 for average temperature, 23.615 for maximum temperature, and 2.16 for minimum temperature reflect the capability of the model to approximate temperatures with a degree of precision that, while varied across different temperature ranges, offers a foundational level of accuracy.

The Max AE demonstrates a significant range, from 33.239 in minimum temperature predictions up to 4034.911 in maximum temperature predictions. While these errors initially may seem high, they also indicate the model's potential to identify and learn from extreme outliers.

The MSE values, particularly the high value observed in maximum temperature predictions, underscore the model's sensitivity to large deviations.

The model shows promising performance in predicting minimum temperatures, as reflected by the relatively lower error metrics, showcasing its strengths in handling at least some aspects of the temperature range.

Alumina cement

Results for testing the model with alumina cement material are presented. Graphical results show that temperature differences are less abrupt that in the soda lime-glass case, and that the model has has much more precision representing the heating curves even if when the difference between the maximum and the minimum are 200ºC.

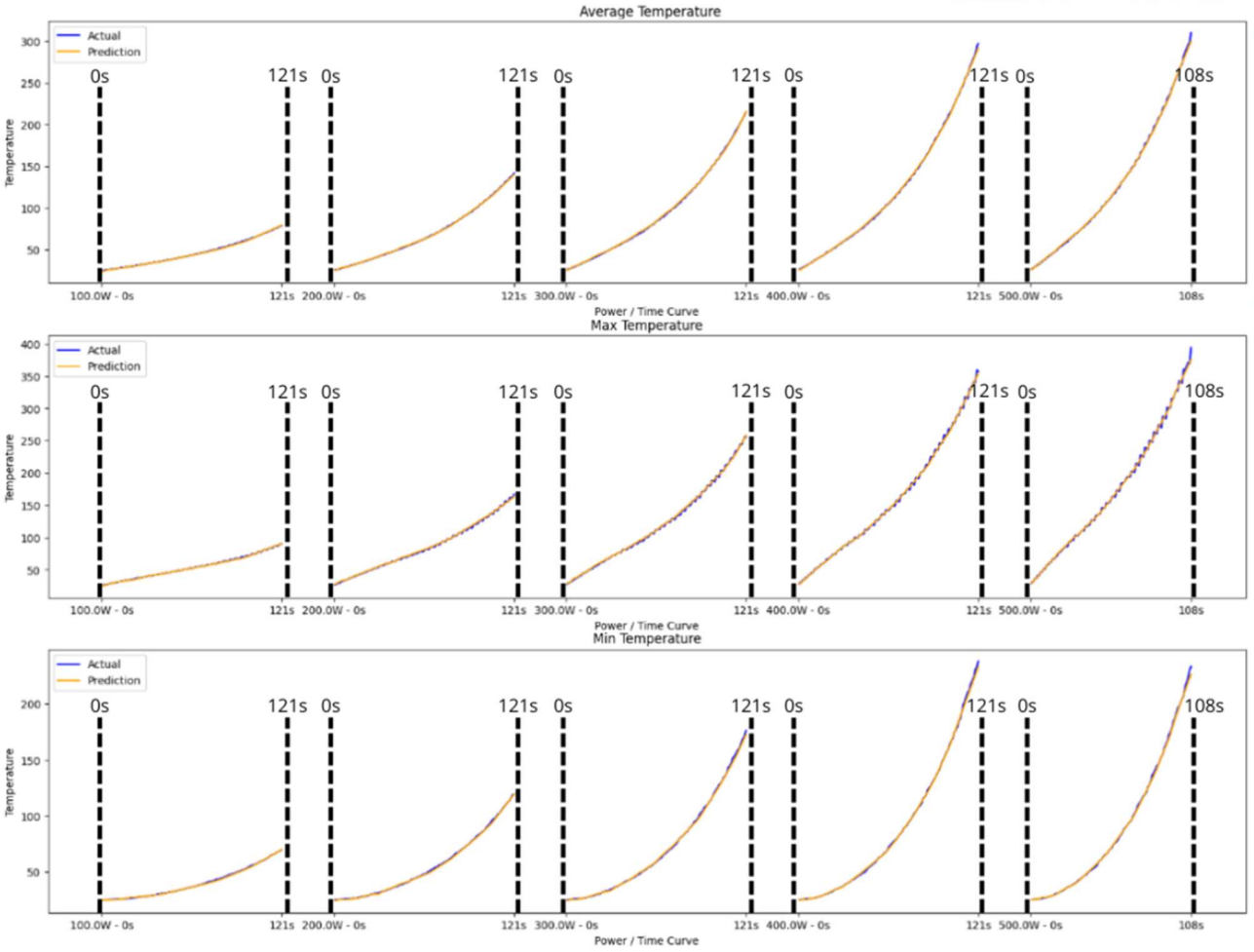


Figure 33. Alumina cement graphical test results

Minimum temperature results: the prediction closely aligns with the actual temperatures, reflecting the model's robust predictive capability for the material's behaviour at varying power levels.

Average temperature results: the predictions show remarkable congruence with the actual temperatures, indicating that the model has effectively captured the relationship between the power applied and the resultant temperature increase.

Maximum temperature results: the model demonstrates an excellent fit with the actual temperature. It captures the trend accurately, although it consistently predicts a marginally lower temperature (<2ºC). This might suggest the model slightly underestimates the peak temperature response of the material or could be accounting for the faster dissipation of heat at higher temperatures.

	Minimum temperature	Average temperature	Maximum temperature
MAE	0,802	1,129	2,475
Max AE	7,165	16,548	37,921
MSE	1,357	2,851	15,098

Table 26. Alumina cement numerical test results

The MAE values of 1.129 for mean temperature, 2.475 for maximum temperature, and 0.802 for minimum temperature showcase the commendable accuracy in predictions. Particularly noteworthy is the precision in predicting minimum temperature, where the error is less than one degree. This level of precision indicates the effectiveness in capturing and predicting the lower range of temperatures with high reliability.

The max AE values, for instance, the maximum error in minimum temperature predictions is significantly lower than that for maximum temperatures, suggesting that the model maintains a tighter control over predictions at the lower end of the temperature spectrum.

The MSE values across the board, with 2.851 for average temperature, 15.098 for maximum temperature, and 1.357 for minimum temperature, point to the overall consistency in performance, with a particularly strong showing in the minimum temperature domain.

The model shows a reasonable level of accuracy with the lowest errors in predicting minimum temperatures and the highest errors for maximum temperatures. The moderate max AE for average temperature and the high MAE error for max temperature indicate that while the model is generally reliable, it does occasionally make significant errors, especially with peak temperatures. The MSE values support this, showing more considerable variability in the errors for max temperatures.

Borosilicate glass

Results obtained when testing model with borosilicate glass are presented. In this case it is evident that, while not significantly changing, the small noise brought about by the maximum temperature curves is somewhat upsetting the prediction.

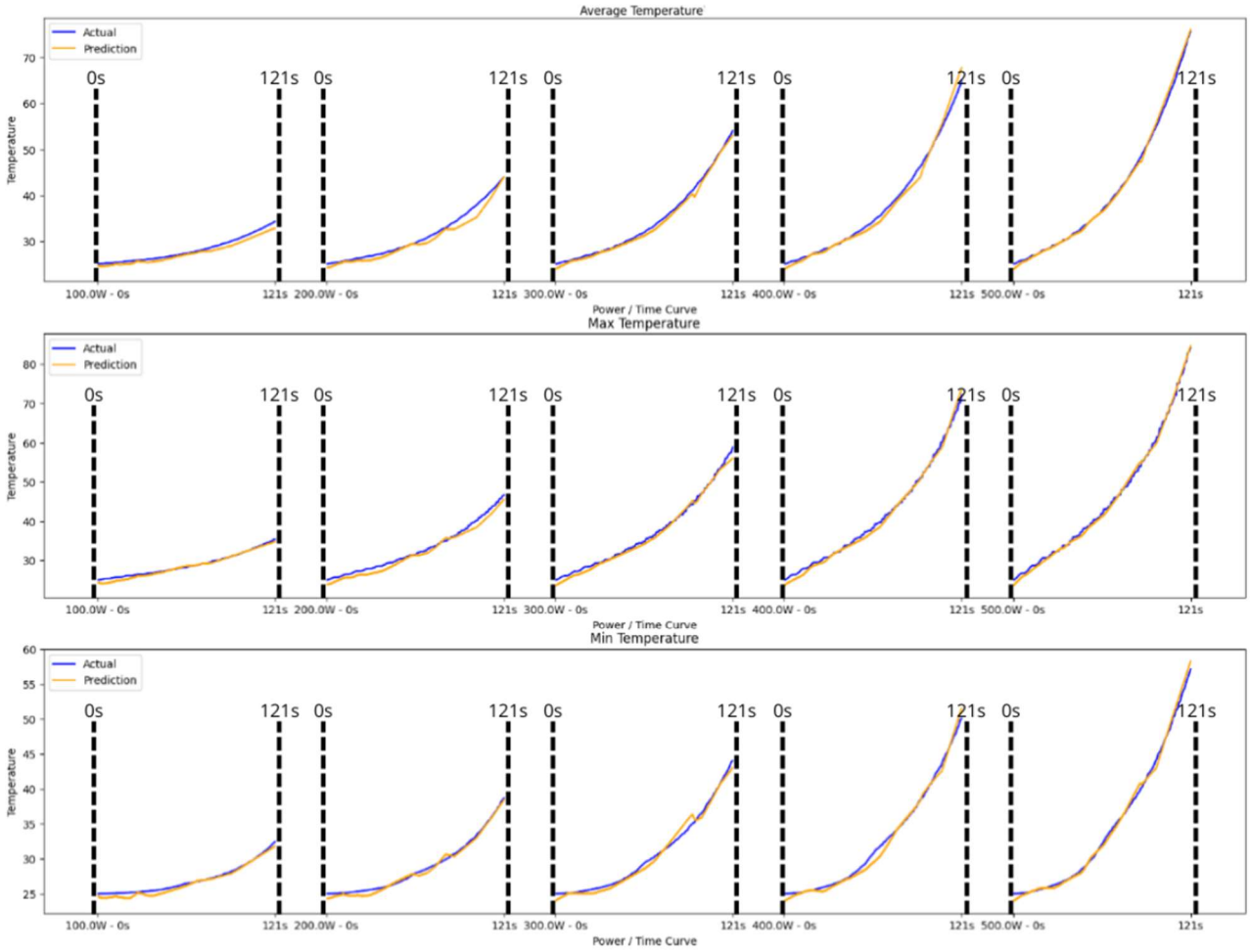


Figure 34. Borosilicate glass graphical test results

Minimum temperature results: predictions remain closely aligned with the actual data, with a very slight underestimation in the early phase of the heating cycle. This tight correlation showcases the robust performance across various power inputs.

Average temperature results: there is an excellent alignment between predicted and actual temperatures, indicating that the model scales its predictions effectively with increased power. The predictions accurately mirror the actual temperatures, showing an understanding of the thermal characteristics across a range of operating conditions.

Max temperature results: the model closely matches the actual temperature profile at these power levels, accurately capturing the response of the material to higher energy inputs. It tracks the rising slope and peak temperatures effectively, with minimal discrepancies.

	Minimum temperature	Average temperature	Maximum temperature
MAE	0,747	0,826	1,073
Max AE	10,680	21,952	34,833
MSE	1,753	2,044	4,442

Table 27. Borosilicate glass numerical test results

With MAE values of 0.826 for average temperature, 1.073 for maximum temperature, and 0.747 for minimum temperature, the model exhibits a commendable level of accuracy across all temperature predictions. These figures are particularly impressive, indicating that, on average, the predictions are very close to the actual temperature measurements.

The max AE values provide insightful indicators of the prediction performance under extreme conditions. The fact that these errors are contained within reasonable limits (21.952 for average temperature, 34.833 for maximum temperature, and 10.680 for minimum temperature) underscores the robustness of the model.

The MSE values reinforce the consistency of the model in performance across different temperature predictions. These metrics, emphasizing the average of the squared differences between predicted and actual values, suggest that the model is effectively capturing the overall temperature trends of the material.

The model demonstrates good predictive performance. However, the relatively high max AE for maximum temperatures indicates that the model may struggle with accurately predicting extreme values. This could be due to several factors, such as the model not being complex enough to capture the nuances of the maximum temperature behaviour or the training data not sufficiently representing the extremes.

Alumina silicate

Now are presented the results obtained during alumina silicate model testing, the last material in the list of semi-transparent materials analysed. Despite of minimal differences between the actual and the predicted minimum temperatures, graphical representation of model testing displays almost null deviations.

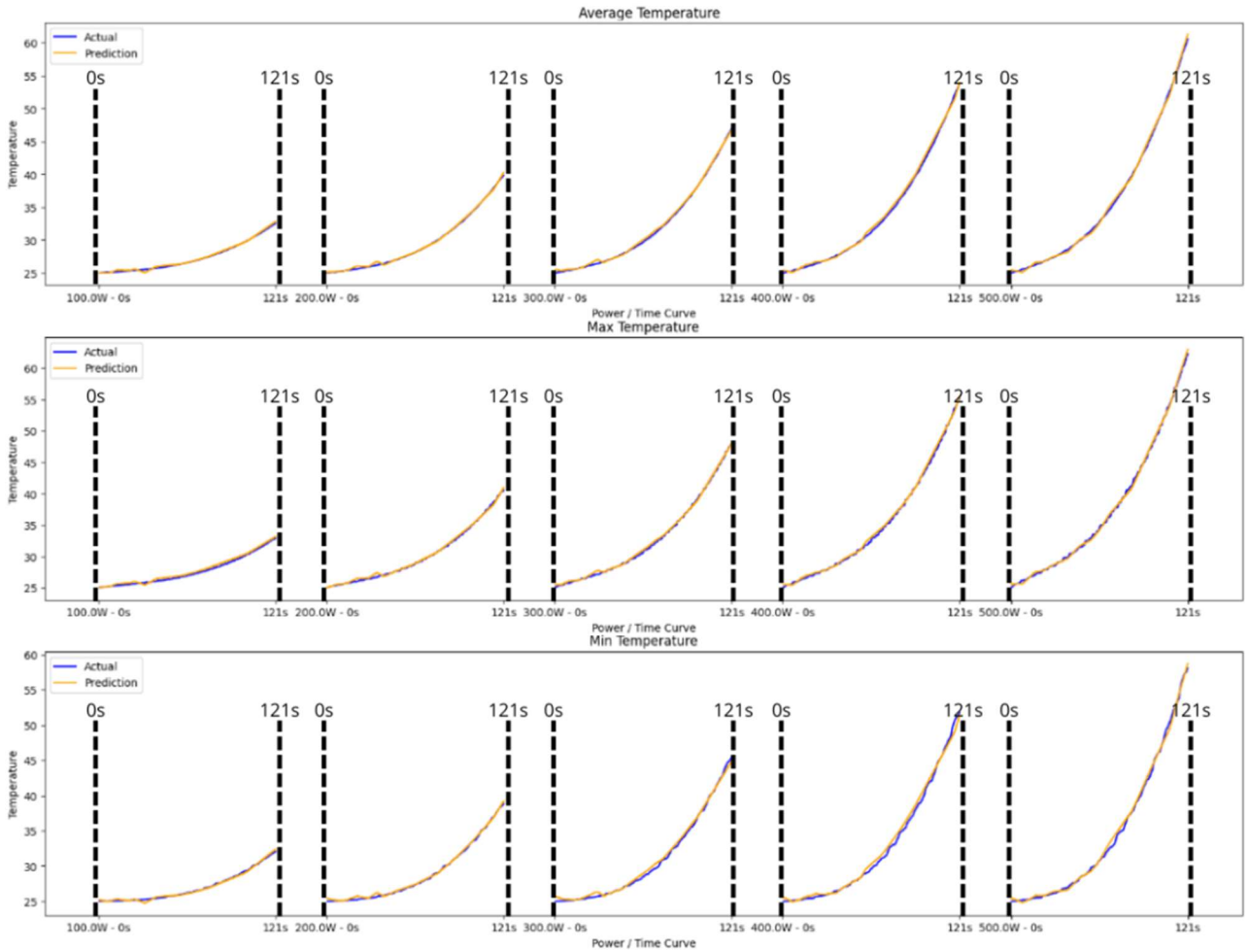


Figure 35. Alumina silicate graphical test results

Minimum temperature results: the prediction is similar to previous ones and strongly follows the actual curve with high fidelity.

Average temperature results: the prediction closely matches the actual temperatures throughout the simulated time, suggesting the model has a solid grasp on the heat absorption of the material and dissipation characteristics under varying power inputs.

Maximum temperature results: the predictions here are also closely aligned, slightly under the actual curve at some points. This could indicate the conservative nature of the model in estimating peak temperatures but still shows a high degree of accuracy.

	Minimum temperature	Average temperature	Maximum temperature
MAE	0,603	0,501	0,567
Max AE	9,848	8,243	8,117
MSE	0,951	0,736	0,865

Table 28. Alumina silicate numerical test results

The MAE values of 0.501 for average temperature, 0.567 for maximum temperature, and 0.603 for minimum temperature are indicative of the high accuracy of the model. Such low MAE figures suggest that, on average, the predictions are very close to the actual temperatures, with minimal deviations.

The max AE values, standing at 8.243 for average temperature, 8.117 for maximum temperature, and 9.848 for minimum temperature, show that the model maintains a good degree of accuracy even in the face of the most challenging predictions. These figures highlight the robustness of the model and its ability to handle outliers or extreme temperature variations without significant deviations from actual values.

With MSE values of 0.736 for average temperature, 0.865 for maximum temperature, and 0.951 for minimum temperature, the model demonstrates consistency in its predictive performance.

These MSE metrics, emphasizing the square of the prediction errors, underscore the effectiveness in accurately forecasting temperatures across a spectrum of conditions with a controlled level of error.

In general, these metrics are indicative of a very well-performing model. The errors are minimal, and even the maximum errors are quite low, which might suggest that the model deals well with outliers or unusual data points.

5.2.3 Susceptor materials

SiC

SiC is the first material studied in the group of susceptors materials. Despite clear non-linearities showed during model testing, graphical representations show good adaptations of predictions to actual values. The model is able to fit predictions to exponential growths showed by material heating curves for each predicted variable. However, the prediction accuracy decreases as the heating power increases due to drastic change in temperature.

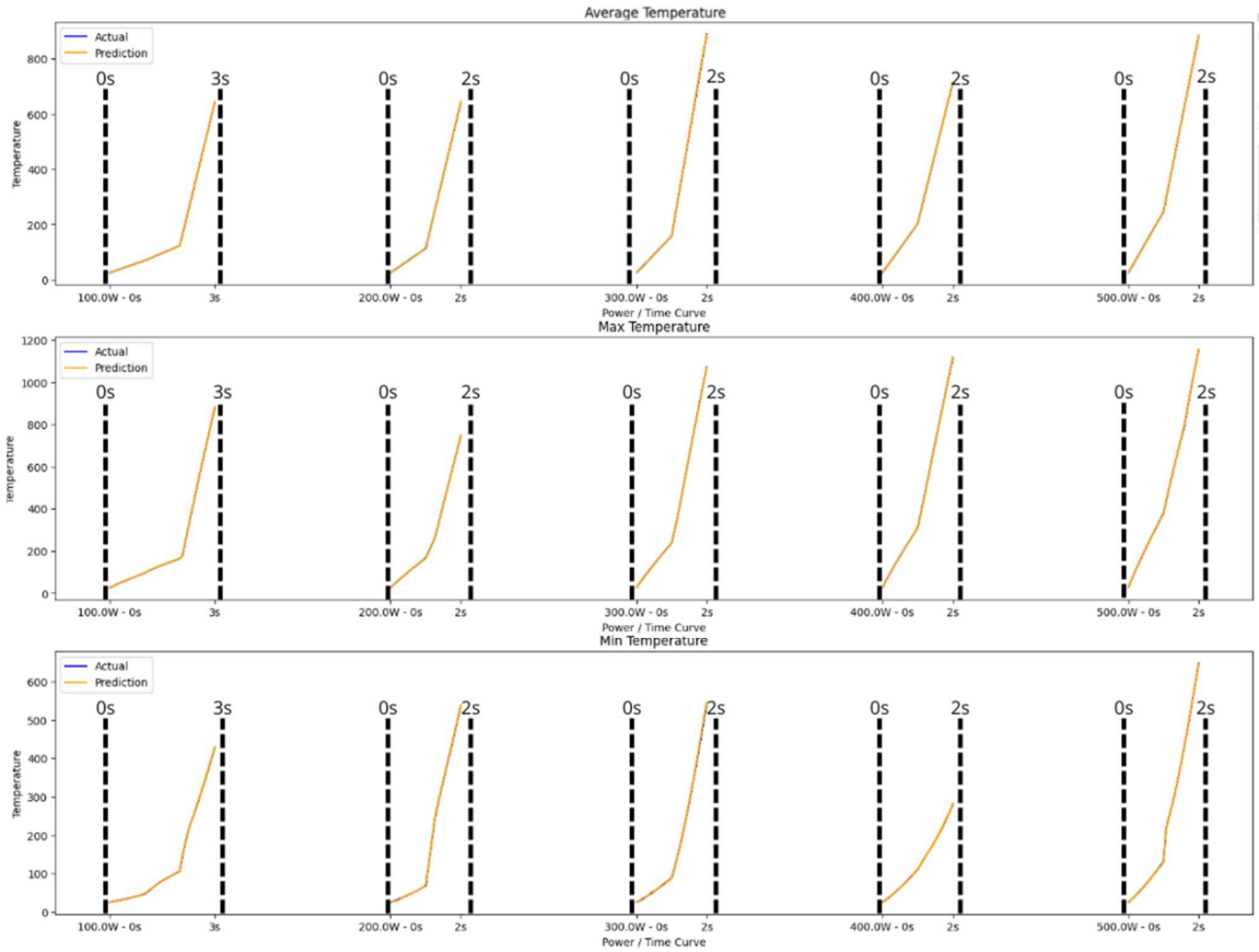


Figure 36. SiC graphical test results

Minimum temperature results: the prediction starts off well-aligned, slightly overestimates, and then follows the actual trend. It captures the initial temperature behaviour accurately.

Average temperature results: the prediction rises sharply and closely follows the actual temperature, indicating a strong initial response. The model captures the rapid increase in temperature, reflecting its understanding of the quick thermal reaction of the material to the applied power.

Maximum temperature results: the predictions at higher power settings exhibit a sharp increase that surpasses the actual temperature initially. This could imply that the model expects a faster thermal response

from the material than what occurs. Despite this, the model predictions rapidly converge with the actual data, indicating the correct overall trend is captured.

	Minimum temperature	Average temperature	Maximum temperature
MAE	1,058	1,190	1,542
Max AE	10,869	12,134	16,706
MSE	2,228	2,781	4,436

Table 29. SiC numerical test results

The MAE values of 1.190 for average temperature, 1.542 for maximum temperature, and 1.058 for minimum temperature demonstrate the commendable accuracy of the model when predicting temperature variations. Particularly, the low MAE for minimum temperature suggests the strong capability in accurately forecasting lower temperature ranges.

While the Max AE values (12.134 for average temperature, 16.706 for maximum temperature, and 10.869 for minimum temperature) highlight the challenges in predicting temperature extremes, they also show the ability to maintain reasonable accuracy under diverse conditions.

The MSE values of 2.781 for average temperature, 4.436 for maximum temperature, and 2.228 for minimum temperature reflect the overall consistency of the model and the accuracy of its predictions. These MSE metrics underscore the effectiveness in capturing the temperature behaviour of the material.

Overall, the model appears to perform well, especially for minimum temperature predictions, with moderate accuracy for mean and maximum temperatures. The higher error metrics for the maximum temperature suggest that the model may not capture extreme values as reliably as it does average or minimum values. This is not uncommon, as extreme values can be more difficult to predict due to their potentially volatile nature.

ALN Powder Compact

The ALN Powder Compact material testing results are shown in the next figure. Overall, graphical results suggest reasonably good performance of the model for average and minimum temperatures.

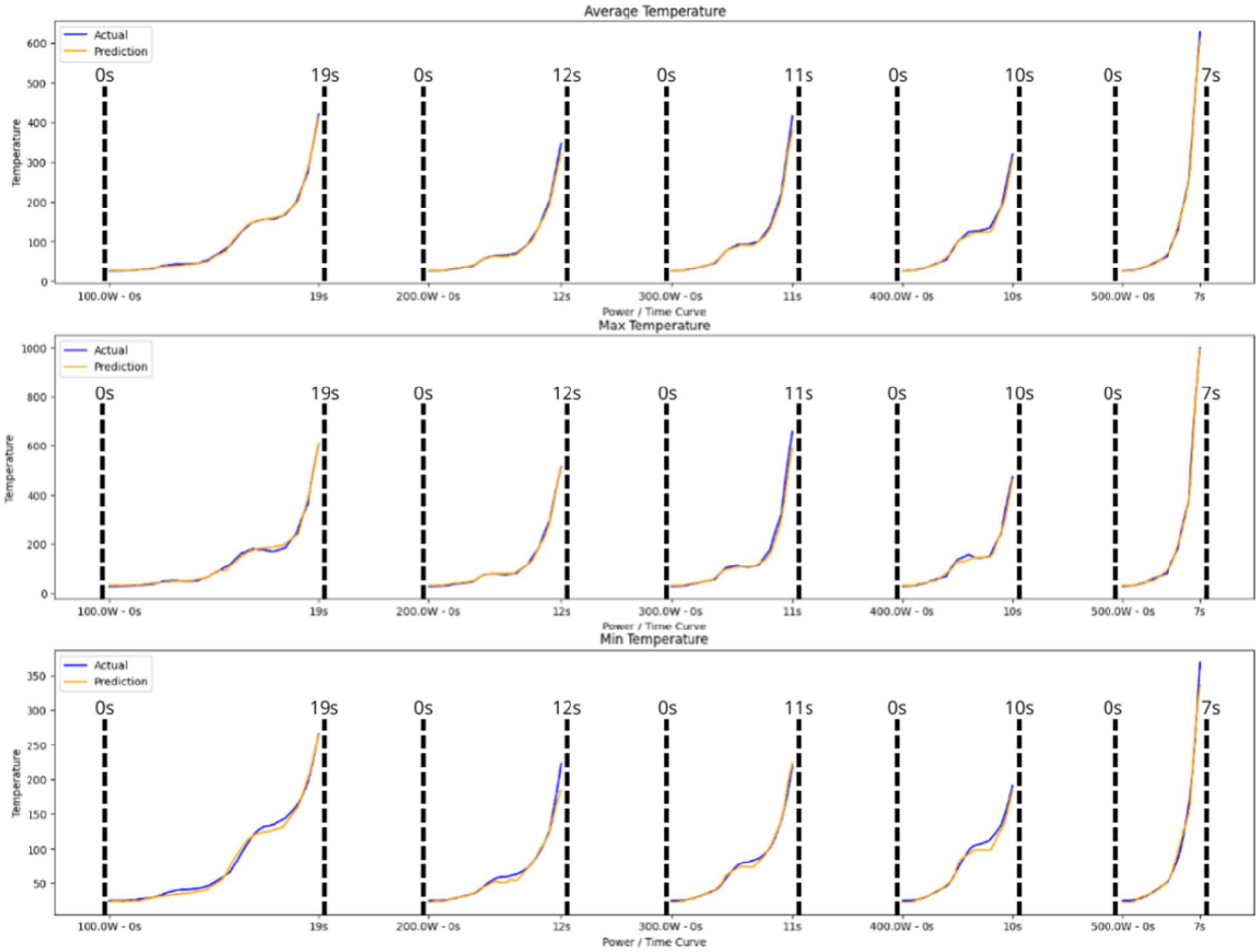


Figure 37. ALN powder compact graphical test results

Minimum temperature results: the model shows a slight initial lag in the prediction, followed by a tight correspondence with the actual temperature, reflecting its capacity to quickly adjust to the thermal response.

Average temperature results: the predictions are impressively close to the actual temperatures, with the model capturing the sharp rise and overall trend accurately. At higher power settings, the prediction and actual curves tightly align, indicating the model effectively adapts to the increased energy input.

Maximum temperature results: there is a very close fit between the predicted and actual temperatures, particularly at the final stage of the heating process, suggesting the model has effectively learned the response of the material at these levels of power input.

	Minimum temperature	Average temperature	Maximum temperature
MAE	3,797	3,353	6,679
Max AE	38,010	47,366	114,272
MSE	29,520	28,107	119,875

Table 30. ALN powder compact numerical test results

The MAE values of 3.353 for average temperature, 6.679 for maximum temperature, and 3.797 for minimum temperature show that the model has a reasonable level of accuracy in its predictions, particularly considering the complex nature of temperature behaviour in this exact material in which the curves have great variations.

Although the max AE values are relatively high, with 47.366 for average temperature, 114.272 for maximum temperature, and 38.01 for minimum temperature, they offer valuable insights into the performance of the model under extreme conditions. These measurements highlight the toughest scenarios where the predictions diverge from actual temperatures.

The MSE values of 28.107 for average temperature, 119.875 for maximum temperature, and 29.520 for minimum temperature point out how the abrupt differences with maximum temperatures are affecting results.

As indicated by the lower MAE and MSE values, the maximum temperature predictions are less accurate, which is common in many predictive models, especially if the maximum temperature data have more volatility or less pattern consistency. The higher max AE values indicate the presence of outlier predictions or instances where the model significantly deviates from the actual values, which might be due to extreme conditions or anomalies in the data that the model has not learned to predict accurately.

CuO

The Cu0 depicted in the next figure. These results have some particularities, and even if the model can adapt itself to the exponential growth when the three curves have different types of growth, the backpropagation of the error from the average and maximum temperature curves leads to the min temperature curve to have some small irregularities at the time of the inference.

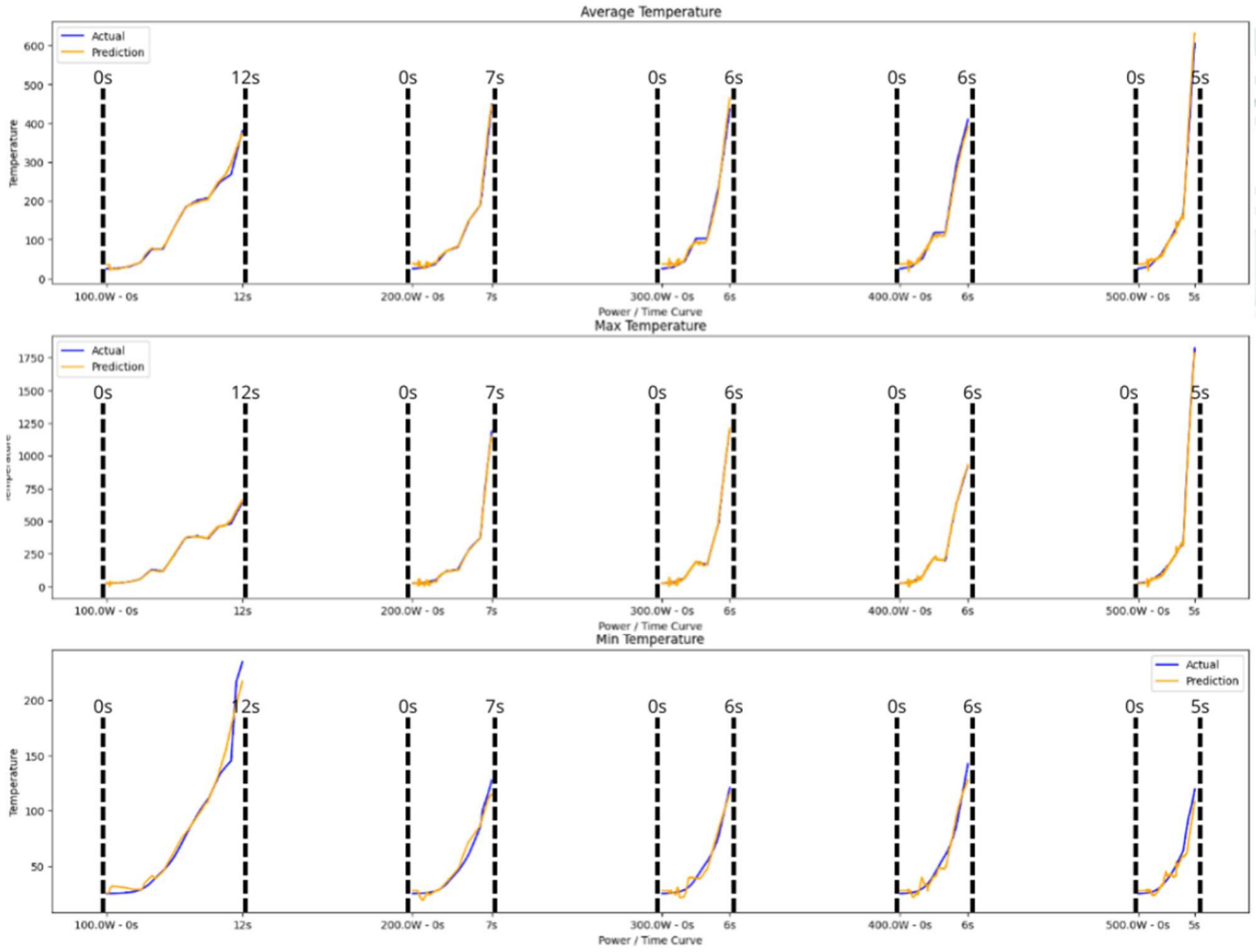


Figure 38. CuO graphical test results

Minimum temperature results: the prediction begins well-aligned but then diverges slightly, indicating the initial accuracy but suggesting a need for adjustment in response to the increasing rate of temperature in material.

Average temperature results: despite a brief period of overestimation, the prediction aligns well with the actual temperature across these power settings. The quick realignment suggests the model adapts efficiently after recognizing its initial prediction error.

Maximum temperature results: a similar pattern to the average temperature, with an initial prediction above the actual temperature, followed by a closer alignment, showing the rapid adjustment capabilities of the model.

	Minimum temperature	Average temperature	Maximum temperature
MAE	4,486	11,978	15,877
Max AE	44,788	208,080	349,268
MSE	46,591	368,710	876,316

Table 31. CuO numerical test results

The MAE values, standing at 11.978 for average temperature, 15.877 for maximum temperature, and notably lower at 4.486 for minimum temperature, indicate a spectrum of accuracy across different temperature predictions. Particularly encouraging is the performance when predicting the minimum temperature with a relatively low error, underscoring the ability of the model to capture lower temperature ranges accurately.

While the max AE values are considerable (208.080 for average temperature, 349.268 for maximum temperature, and 44.788 for minimum temperature), they provide a clear benchmark for the performance under the most challenging conditions. The significantly lower max AE for minimum temperature further accentuates the relative prediction strength in this area.

The MSE values (368.710 for average temperature, 876.316 for maximum temperature, and 46.591 for minimum temperature) reveal areas where the predictive accuracy can be substantially improved. The relatively lower MSE for minimum temperature predictions again highlights this as a strength of the current model, providing a foundation for building upon and extending this accuracy across other temperature ranges.

The metrics suggest that the model is most accurate with the minimum temperature predictions and least accurate with the maximum temperature predictions. The high max AE in both average and maximum temperatures could indicate outliers or instances where the model fails to capture extreme values accurately. Differences in these metrics among the three temperature categories may suggest that different aspects of the model feature extraction or learning process are better suited to capture certain types of temperature behaviour than others.

7 Conclusions

This work highlights the integration of ML with materials science, demonstrating the potential for significant advancements through this interdisciplinary approach. The focus of this research is twofold: bio-based materials from the GREEN LOOP project and materials used in microwave heating processes due to their specific thermal properties.

In materials science, the characterization of material properties is crucial for understanding behaviour and performance. By leveraging data science and ML techniques, it is possible to derive valuable insights from material data, enabling the discovery of new opportunities for optimizing, designing, and studying materials. The results of this work showcase the versatility of ML in predicting a wide range of material properties, including tensile strength, elastic modulus, material elongation, compression, and thermal response to microwave excitation.

The applicability of ML in this context extends beyond just mechanical properties. It can also be effectively employed to analyse thermal, electrical, and chemical behaviours. This demonstrates the broad relevance of ML to various aspects of materials science. Given the flexibility of ML as a field, the selection of appropriate techniques depends on the quality and quantity of the data available, as well as the specific research objectives. This data-driven approach is fundamental, as the success of ML applications in materials science is highly dependent on the robustness and accuracy of the underlying data.

ML algorithms are inherently versatile, and their application is not constrained by the specific material being analysed. This is because ML algorithms process data, which is a representation of real-world phenomena, regardless of its origin or meaning. In this context, the models demonstrated strong performance across a variety of materials, showcasing their adaptability to different material types and properties, and essentially demonstrating that they were able to uncover meaningful relationships and trends from data.

As indicated in the literature, there is no universal consensus on the most appropriate algorithm for a given task in ML. Instead, algorithm selection is part of a nuanced, iterative process that involves aligning the capabilities of various algorithms with the specific goals and characteristics of the project. This informed exploration requires a deep understanding of both the data and the task at hand. In this case, from this exploration, a subset of candidate algorithms is identified, which are then rigorously tested, evaluated, and refined based on their performance metrics. Specifically in this work, ANN, KNN and CNN have been studied and developed for the specific problems managed, showing good performances for the assigned tasks.

This underscores the notion that ML is as much an art as it is a science. The path to optimal performance is not rigidly defined. Instead, success often hinges on key decisions made throughout the project lifecycle, from data preprocessing to model selection, to tuning architectures and hyperparameters. Each of these decisions can significantly influence the outcomes, and potential improvements often emerge from creative experimentation and refinement at each stage. For this reason, the iterative and experimental nature of ML underscores its flexibility and the need for a tailored approach in every unique context.

8 References

[1] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks", Accessed: Aug. 06, 2024. [Online]. Available: http://www.iro.umontreal.

[2] Naveen Kumar Thawait and Dr. Umakant Shrivastava, "Machine Learning Techniques for Predicting Conductive Properties of New Materials," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 10, no. 3, pp. 576–585, Jun. 2024, doi: 10.32628/cseit2410340.

[3] R. A. Matula, "The Importance of Numeric Databases to Materials Science."

[4] O. TA, "Application of Machine Learning Tools to Material Science: A Mini-Review," *Physical Science & Biophysics Journal*, vol. 8, no. 1, pp. 1–6, Feb. 2024, doi: 10.23880/psbj-16000269.

[5] S. Gong, "The significance of materials informatics on material science," *Applied and Computational Engineering*, vol. 58, no. 1, pp. 208–214, Apr. 2024, doi: 10.54254/2755-2721/58/20240723.

[6] D. Merayo, A. Rodríguez-Prieto, and A. M. Camacho, "Topological optimization of artificial neural networks to estimate mechanical properties in metal forming using machine learning," *Metals (Basel)*, vol. 11, no. 8, Aug. 2021, doi: 10.3390/met11081289.

[7] M. Raparthi *et al.*, "Investigating the use of Deep Learning, in Materials Research for Predicting Material Properties, Identifying new Materials, and Optimizing Material Selection for Mechanical Components Assistant Professor, Kalasalingam Academy of Research and Education," 2024. [Online]. Available: https://ballisticsjournal.com

[8] S. Wang, T. Zhu, and J. Grossman, "Multimodal machine learning for materials science: composition-structure bimodal learning for experimentally measured properties." [Online]. Available: https://www.researchgate.net/publication/373837037

[9] A. Suryawanshi and N. Behera, "Prediction of mechanical properties of dental composite materials using machine learning algorithms," *Materwiss Werksttech*, vol. 54, no. 11, pp. 1350–1361, Nov. 2023, doi: 10.1002/mawe.202200294.

[10] K. Murad Ali, "Enhancing Material Property Predictions through Optimized KNN Imputation and Deep Neural Network Modeling," *IgMin Research*, vol. 2, no. 6, pp. 425–431, Jun. 2024, doi: 10.61927/igmin197.

[11] Sam. Lau, Joseph. Gonzalez, and D. Ann. Nolan, *Learning data science : data wrangling, exploration, visualization, and modeling with Python*. O'Reilly Media, Inc., 2023.

[12] "Data Mining Third Edition."

[13] "Nicolas Vandeput Data Science for Supply Chain Forecasting."

[14] A. Zheng and A. Casari, "Feature Engineering for Machine Learning PRINCIPLES AND TECHNIQUES FOR DATA SCIENTISTS."

[15] K. M. Sujon, R. Binti Hassan, Z. T. Towshi, M. A. Othman, A. Samad, and K. Choi, "When to Use Standardization and Normalization: Empirical Evidence from Machine Learning Models and XAI", doi: 10.1109/ACCESS.2017.DOI.

[16] "KerasTuner." Accessed: Sep. 25, 2024. [Online]. Available: https://keras.io/keras_tuner/

[17] "Keras: Deep Learning for humans." Accessed: Sep. 25, 2024. [Online]. Available: https://keras.io/

[18] S. C. Bhatt and N. D. Ghetiya, "3D Multiphysics simulation of microwave heating of bulk metals with parametric variations," *Chemical Engineering and Processing - Process Intensification*, vol. 184, Feb. 2023, doi: 10.1016/j.cep.2023.109271.

[19] R. Yang and J. Chen, "Mechanistic and Machine Learning Modeling of Microwave Heating Process in Domestic Ovens: A Review," *Foods*, vol. 10, no. 9, Sep. 2021, doi: 10.3390/FOODS10092029.

[20] R. Yang, Z. Wang, and J. Chen, "An Integrated Approach of Mechanistic-Modeling and Machine-Learning for Thickness Optimization of Frozen Microwaveable Foods," *Foods 2021, Vol. 10, Page 763*, vol. 10, no. 4, p. 763, Apr. 2021, doi: 10.3390/FOODS10040763.

[21] B. Milovanovic, "Microwave applicators modeling Alternative approaches based on neural networks incorporating domain knowledge," *… Network Applications in …*, Accessed: Oct. 03, 2024. [Online]. Available: https://www.academia.edu/67897208/Microwave_applicators_modeling_Alternative_approaches_based_on_neural_networks_incorporating_domain_knowledge

[22] C. Lewis, J. Bryan, N. Schwartz, J. Hale, K. Fanning, and J. S. Colton, "Machine Learning to Predict Quasi TE Mode Resonances in Double-Stacked Dielectric Cavities," *IEEE Trans Microw Theory Tech*, vol. 70, no. 4, pp. 2135–2146, Apr. 2022, doi: 10.1109/TMTT.2022.3145357.

[23] L. Acevedo, G. Ferreira, and A. M. López-Sabirón, "Exergy transfer principles of microwavable materials under electromagnetic effects," *Mater Today Commun*, vol. 27, p. 102313, Jun. 2021, doi: 10.1016/J.MTCOMM.2021.102313.