REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

Ministère de l’Enseignement Supérieure et de la Recherche Scientifique

Université El-Hadj Lakhdar – Batna 1

Faculté de Science de la Matière

Département de Physique

THÈSE

Présentée en vue de l’obtention du

Diplôme de Doctorat de troisième cycle

par:

Mohamed Mammeri

Thème:

Application des techniques de Machine Learning pour les dispositifs Photovoltaïques et Optoélectroniques

Domaine	: Science de la Matière
Filière	: Physique
Spécialité	: Physique des Matériaux et énergies renouvelables

<u>Devant le jury:</u>

Président:	Belbacha Eldjemai	Prof	Universite Batna 1
Rapporteur:	Dehimi Lakhdar	Prof	Universite Batna 1
Examinateurs:	Guezouli Larbi	Prof	ENSEREDD Batna
	Bendib Toufik	M.C.A	ENSEREDD Batna
Invité:	Bencherif Hicham	M.C.B	ENSEREDD Batna

Democratic and Popular Republic of Algeria

Ministry of Higher Education and Scientific Research

University El-Hadj Lakhdar – Batna 1

Faculty of Science of Matter

Department of Physics

THESE

Submitted in order to obtain

Doctorate Diploma

By:

Mohamed Mammeri

Theme:

Machine learning techniques applications for Photovoltaic and Optoelectronic devices

Domain	: Science of Matter
Sector	: Physics
Title of Training	: Material physics and renewable energy

<u>In front of the jury:</u>

President:	Belbacha Eldjemai	Prof	Universite Batna 1
Supervisor:	Dehimi Lakhdar	Prof	Universite Batna 1
Examiners:	Guezouli Larbi	Prof	ENSEREDD Batna
	Bendib Toufik	M.C.A	ENSEREDD Batna
Guest:	Bencherif Hicham	M.C.B	ENSEREDD Batna

DEDICATION

To my family and my friends.

ACKNOWLEDGEMNTS

All thanks are to ALLAH for His guidance and support, and for granting me the opportunity to accomplish this work. Then, I would thank my supervisors, colleagues, and friends for the guidance and support provided.

I sincerely thank my guide and supervisor Professor Dhimi Lakhdar for taking me in as a graduate student, for giving me the opportunities to pursue my evolving interests in research, and for supporting me throughout all the difficult times I encountered.

I express my sincere thanks to Dr. Bencherif Hichem for working closely with me to understand the results of our research, for exchanging ideas, and for always looking out for new ideas and methods.

I express my sincere thanks to Pr. Fortunato Pezzimenti, Pr. Mongi Amami, and M. Khalid Hossain for working together on various projects, for sharing their knowledge with me, and for supporting my pursuit of the degree.

I express my gratitude to my defense committee, who generously provided knowledge and expertise. Additionally, this endeavor would not have been possible without the generous support from the Laboratory of Physical and Chemical Studies of Materials (LEPCM).

Finally, I am grateful to all of my family members for their continuous support at every stage of my studies.

Table of Contents

List of Tables

List of Figures

Introduction

Motivation for Material Science

Nowadays, human civilization is practically entirely dependent on technological advances in different disciplines, and searching strongly for energy from different sources to accommodate this development. Therefore, it is essential to accelerate the device's development for both technological and energy purposes. For example, accelerating the development of optoelectronic devices improves medical equipment, control access systems, and telecommunications fields like high-speed internet and 5G. Moreover, the progress in the development of materials for photovoltaic device components creates opportunities to obtain safe, clean, and renewable energy. In fact, there is a significant growth in solar energy production from less than 1TWh in the 1990s to more than 3400TWh in 2022 (see Figure 1) [1]. This progress was accompanied by the discovery of new generations of solar cells, and the development of new materials to achieve stable and high-performance solar cells.

In general, the material development process has been passed through different paradigms over history (see Figure 2) [2]. The first paradigm was based on trial-and-error experimentation, this classic approach relied on giving a solution based on the experience and the intuition of the material scientist, then learning from the failure and trying again [3]. This approach is extremely time-consuming and exhausting the resources. For instance, Thomas Edison discovered the carbonized cotton used in the light bulb after many failed experiments [4].

In the second paradigm, the materials scientist used the physical laws and semi-empirical models gathered from the experimental results to lead the discovery of new materials. In the late

20th century, the development of powerful computers led to the computing of new materials proprieties by using first principles and solving the Schrodinger equation. For example, the *ab initio* simulation method, and finding the material proprieties using Density Functional Theory (DFT). The success of this approach led to the establishment of many aspiring projects. For example, the Material Project [5], the Material Genome Initiative (MGI) [6], and the Open Quantum Materials Database (OQMD) [7]. Additional materials databases and projects are summarized in Ref. [8-14]. These projects provide very large computation data for the design of new materials. However, given the fact that these simulation methods are powerful and were used to discover new materials, but the very expensive computational costs and sometimes required large experimental data (e.g. Computational thermodynamics) posed a significant limitation.

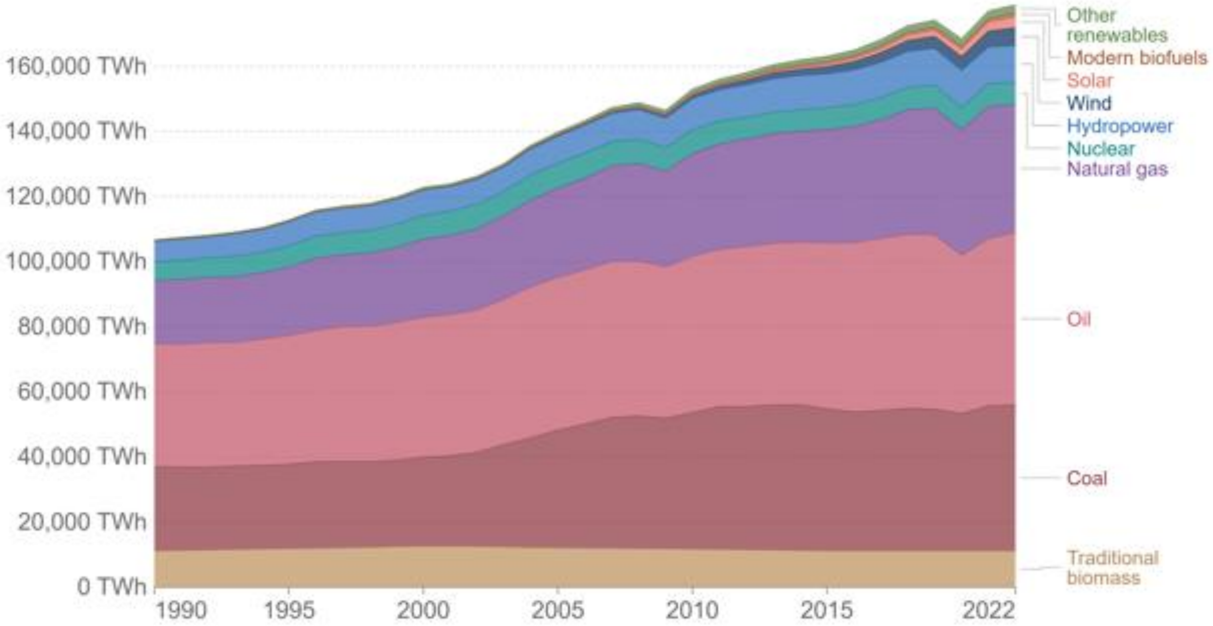


Figure 1- Global primary energy consumption by source

The fourth paradigm is guided by data-driven approaches like Machine learning (ML), Data Mining, etc. Recently, the development of the ML field has provided a tremendous revolution in many disciplines including materials design. The main idea of the ML approach is to create a model derived from experimental and/or computational data when the analytic models are not promising, this model reflects the patterns and the relationships in the data and is used to guide the design of new materials. In particular, the main advantage of this approach is it usually much faster than the simulation and experiments approaches, and can extract highly complex patterns from the data. However, there are several challenges associated with extracting knowledge from large volumes of data. In many cases, the accuracy of the ML model is inefficient enough to meet the material design needs. However, it is still in the early stages of development.

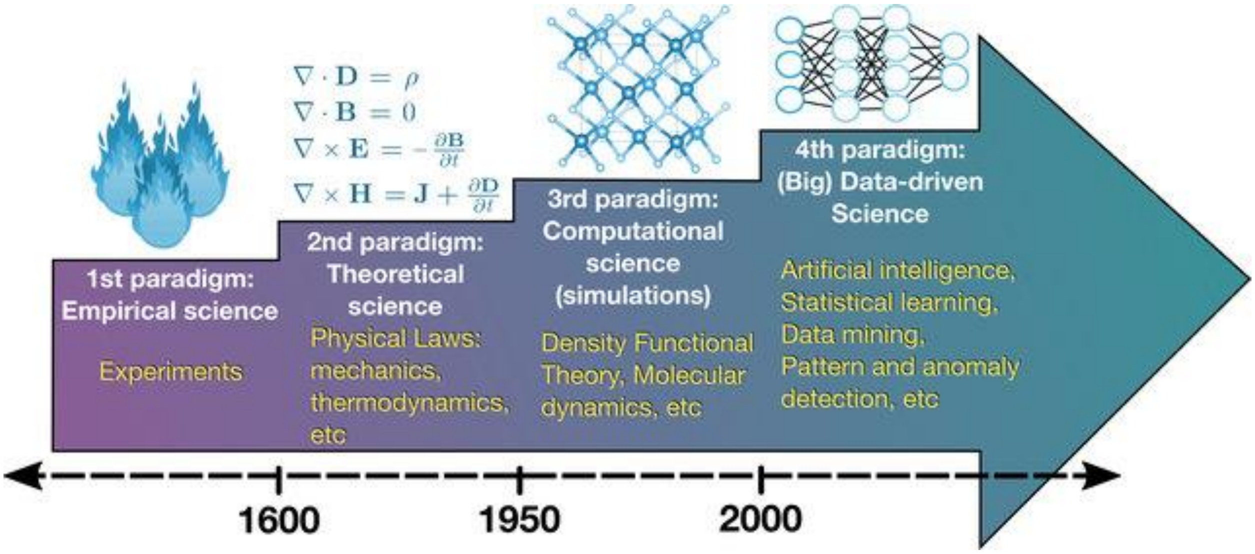

$\nabla \cdot \mathbf{D} = \rho$
$\nabla \cdot \mathbf{B} = 0$
$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}$
$\nabla \times \mathbf{H} = \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t}$
1st paradigm: Empirical science
Experiments
2nd paradigm: Theoretical science
Physical Laws: mechanics, thermodynamics, etc
3rd paradigm: Computational science (simulations)
Density Functional Theory, Molecular dynamics, etc
4th paradigm: (Big) Data-driven Science
Artificial intelligence, Statistical learning, Data mining, Pattern and anomaly detection, etc
1600
1950
2000

Figure 2- Four paradigms of material science [15].

Problem statement and Overview of the proposed approaches

The Perovskite Solar Cells (PSC) is a third-generation solar cell that elicited the interest of many researchers. Since their discovery in 2009, the PSCs attained a massive boost in their power conversion efficiency (PCE) from 3.8% in 2009 to 26% in 2023 [16]. Along with their easy and cheap fabrication cost, these two factors place the PSCs as an important competitor in solar energy generation. However, the barrier to the successful commercialization of this technology is principally due to its instability. In fact, stability, efficiency, and costs are the golden triangle for a successful practical application. Therefore, there is a progressively continuous shift in research toward the investigation of PSC stability. Figure 3 shows a simple research in Science Direct by using the terms of Perovskite Solar Cell and stability (using AND operator), in 2022 the number of papers involving the stability of the PSC reached 3438 paper compared to 74 paper in 2010s, which indicate a great interest in carried out this topic. Consistently, these researches generate a huge amount of accumulated experimental and computational data. However, earlier the data collected were neglected and many opportunities to improve the understanding of the PSC devices were missed because of the lack of efficient methods for addressing these data. Therefore, in this work, we have used ML techniques to extract the general patterns and important and useful information from the PSC device data.

In this thesis, our objective is to use different ML and Artificial Neural Networks (ANN) techniques to provide clear guidance to solve different problems that hinder the development of a practical PSC device. In each chapter of this work, we will focus on one specific problem. Which will be introduced at the start of the chapter. Also, each chapter contains a detailed definition and description of the ML technique used. Then, provide the details of the methods and materials used. And finish with an application of the proposed approach to solve the assigned problem. The problems that will be discussed in this thesis are as follows:

Problem 1: *Investigating different ML and ANN algorithms to find the most suitable algorithm for our dataset.* In Chapter 1, we provided a very simple introduction to the basic concepts of machine learning in general and a brief explanation of how the techniques used in this research works, then in the further chapters we will dive into more depth of these techniques. In Chapter 2, we present a general idea of how to apply different machine learning techniques to our dataset and compare them to extract the most suitable one for our problems and determine the strategy to improve the performance of the ML models.

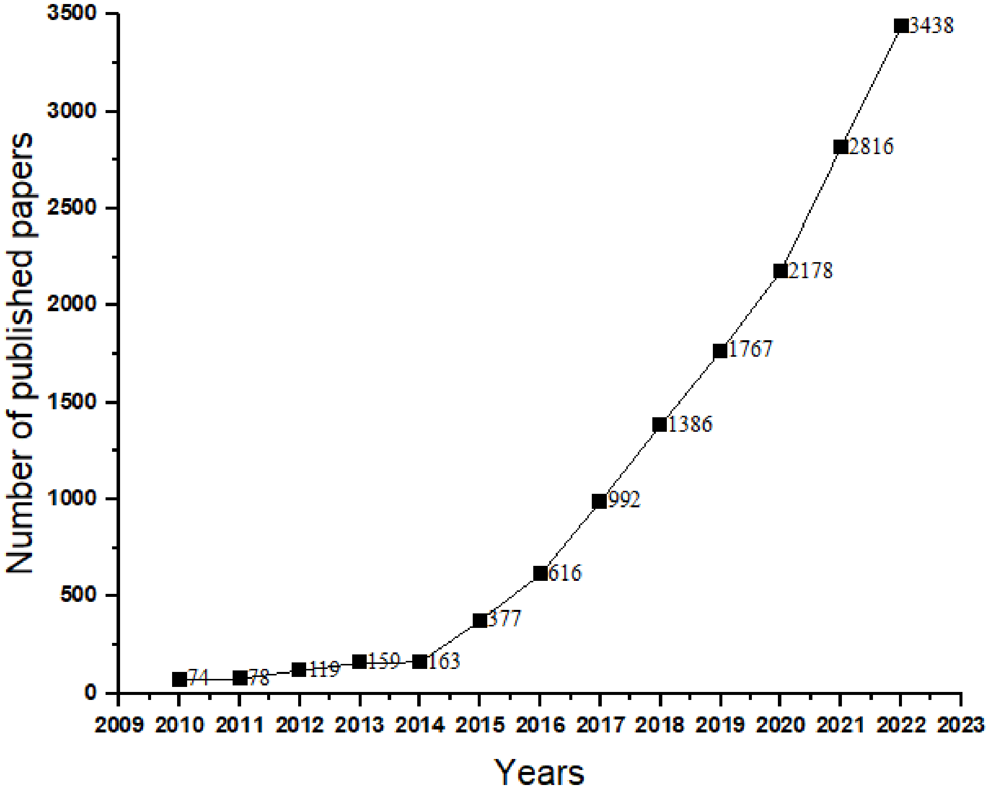


Figure 3- Number of papers published involving PSC stability per year

Problem 2: *Investigating the different factors that influence the degradation of the PSC device's stability.* In Chapter 3, we used Extreme Gradient Boosting and MultiLayer Perceptron algorithms to investigate the influence of the back contact on the total device stability and estimate the adequate material that reduces the device stability degradation. In Chapter 4, we explore a large number of different extrinsic and intrinsic factors and analyze their influence on PSC device stability by using the Extra Tree algorithm. The key motivation of this chapter is to extract the important factors that cause the device instability and propose various solutions to achieve long-term stability.

Problem 3: *A guide to improve the PSC efficiency through optimizing the device layer proprieties and finding adequate materials for this p*urpose. In Chapter 5, we focused on enhancing the PSC device's power conversion efficiency using the Random Forest algorithm. The approach employed is based on finding the optimum device proprieties like band gap and the thicknesses of the active layer and recommending candidate materials for each layer that are anticipated to enhance the device PCE.

Chapter 1. Simple Introduction to Machine Learning

1.1 Introduction

The Machine learning field has generated eminently attention in recent years due to the fact that it has demonstrated significant capabilities in different tasks. However, the ML field is a relatively old field that has been studied for decades. Since W.Pitts and W. McCulloch published a paper in 1943 involving a mathematical modeling of Neural Networks and decision-making in cognitive systems [17]. Thereupon, the idea of Artificial Intelligence growth. In 1950 a tuning test to determine computing intelligence was created by Alan Turing [18]. Where the first drafting of the term "Machine learning" was in 1959 by Arthur Samuel in his paper entitled "*Some studies in machine learning using the game of checkers*" [19]. In 1962 he lost in that game (checkers) against the computer which was considered at that time a big milestone in this field. From this point, machine learning techniques and computer programs start to develop. After the early 1980s, the first real-world applications of ML appeared [20]. Interestingly, the two main conferences on machine learning started at that time, the International Conference on Machine Learning (ICML) in 1980 and Neural Information Processing Systems (NeurIPS, formerly NIPS) in 1987 [20]. In 1989, an early remarkable achievement has developed from machine learning: spoken word recognition [21], and the autonomous driving car [22]. And in 1997, Deep Blue (the famous chess-playing system) shocked the world by beating the world champion in the chess game [23]. Figure 1.1 Cite a few historical landmarks that are significant to the development of ML. However, over the past years, the ML field has grown tremendously and started to be widely used. From daily life like movie recommendations to industry, medicine, and public health. So what is ML? And what is the basic attribute of ML?

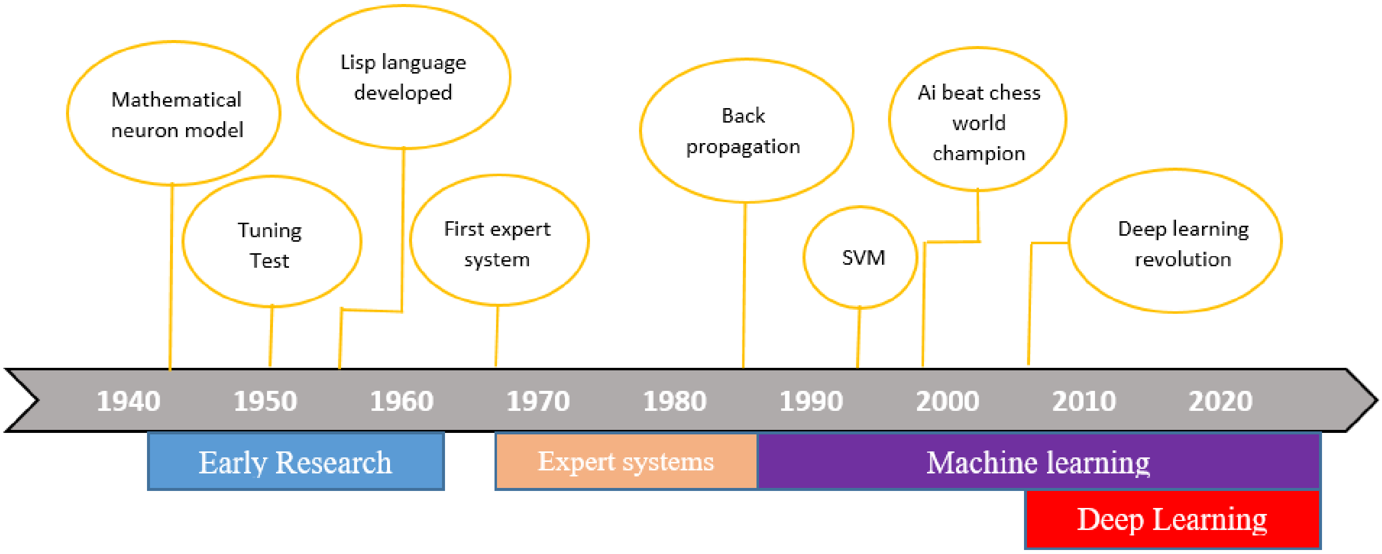


Figure 1.1- Some of the landmarks of ML development

Machine learning is considered as a branch of Artificial Intelligence (AI). Where a system can learn and improve based upon data without being explicitly programmed by using algorithms that can imitate the way humans learn [19][24]. ML models can be classified into three types depending on the method of training: Supervised learning, Unsupervised learning, and Reinforcement learning.

Before we go further into details, let's see in brief the machine learning process. So, one of the key things is that we are going to give machine examples. And these examples are characterized by data samples consisting of inputs and outputs (target). The ML techniques analyze the data and - by itself – create a "model" out of it. Then, this model is used to achieve what we want as the final product [25]. For instance, ML models can be used for forecasting, Object Detection, Pattern Recognition, Cluster analysis, and more. Figure 1.2 illustrates the ML process.

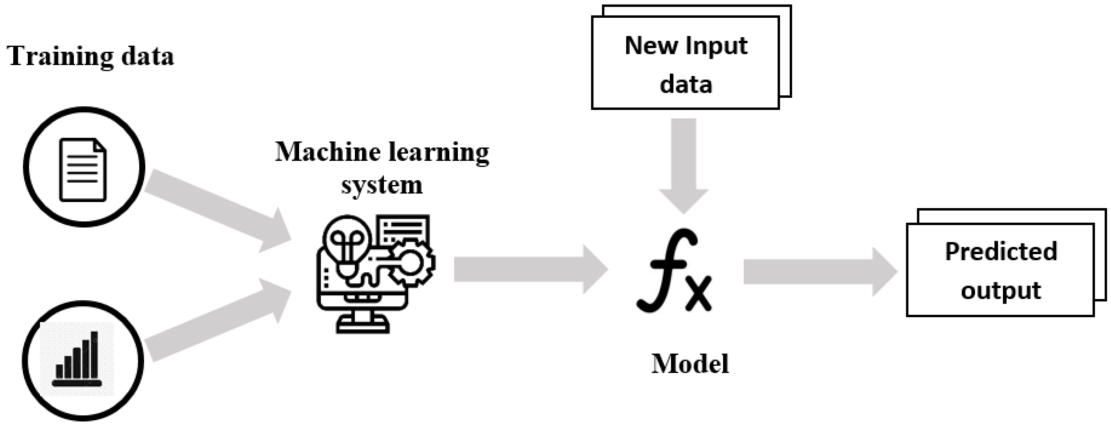


Figure 1.2- What happens in the ML process

In particular, if the training data consists of inputs and output pairs {inputs, output}, and the ML aims to learn a mapping from the inputs to an output, this method is called Supervised Learning, where the correct output is provided by a human (supervisor) during the training process. In Unsupervised learning, the data consists only of inputs {input}, and there is no supervisor [24]. Generally, unsupervised learning is used for investigating and finding the regularities in the data. Reinforcement learning is employed when there is a sequence of actions to reach the goal. Generally used to generate a policy when optimal interaction is required [24, 25]. In this research, we will focus on supervised learning where the inputs are all aspects of Perovskite Solar cell (PSV) manufacturing and the output is either power conversion efficiency (PCE) or long-term stability. The reason for choosing PSV will be discussed further in the next chapter.

Before we jump into the practical part. We must illustrate - to some extent- the main concepts behind some ML techniques used in this work. Understanding the specificities of ML techniques is vital for choosing the most optimal inference approach and foundational for understanding the next chapters.

1.2 Machine learning Techniques

1.2.1 Regression and Classification

Regression and Classification are the typical application types that are frequently used in supervised learning. In machine learning, Classification refers to a problem with predictive modeling where the model finds the class to which the data belongs [25]. Some examples of classification problems:

Given an Email → Classify if it is Spam or Regular

Given a Document → Classify the genre of the document (science, sport, ….etc)

Face recognition services → classify if one of the registered users

Taken the first example, the Email is the input for the model which will predict to which class it belongs. Spam class or Regular class. In contrast, the Regression model will estimate a continuous value of the output instead of determining a specific class. As an example of regression, forecasting the weather given related features like temperature, humidity … etc. The different input factors are called "features". Another example of regression –from this research- is the estimation of the PSV PCE given the cell configuration materials. In summary, the ML analysis provides classification when we want the model to determine in which group of data the input belongs, and regression when we want the model to estimate required values.

1.2.2 Overview of the machine learning process

The objective of this section is to illustrate the learning processes that are associated with most supervised learning methods. The task of machine learning is to create a model which will provide an output from a given input. The input is denoted as x, and the output is denoted as y. the nature of x can take different forms. It can for example be a single number, vector or matrix, and even image or text. Typically, in our case, the nature of y takes two forms. A real number in the case of regression, or it corresponds to a label in the case of classification.

Actually, the ML model can be considered as a function that transforms x into y. So that $y = f(x)$. The function f also can be artificial neural networks. Wherein the form of f is as follows:

$$f(x) = W.x + b \quad (1)$$

Where W is called weight and b bias which represent the learnable parameters, W controls the signal (strength) of an input feature, and b is an additional unit that helps to correct the model. Moreover, W and x could be a matrix where the columns in x represent a training example and the rows represent a specific feature, while W represents the corresponding weights.

Consider another function ℓ called the Loss function that measures the error between the output of the function f and the correct output provided by the supervisor. The closer the loss function is to zero, the closer the output of f is to the true value. There are several forms of Loss function. For example, the classical least square error function. That is such that

$$\ell(f(x), y) = (f(x) - y)^2 \ . \quad (2)$$

The data provided for the learning process is called the Training set and is denoted (x_n, y_n). Where n represents the n^{th} example of the training set. However, the training process involves finding the function f that minimizes the average error of the training set which is called the Cost function and denotes J. Which satisfies:

$$J(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(x)^i, y^i) \quad (3)$$

The procedure of minimizing the cost function is called *the optimization procedure.* Usually, the function f is iteratively modified (by modifying w and b) until the cost function reaches the global minimum [20]. However, minimizing the cost function doesn't guarantee that the model succeeds in the generalization for new data, it usually happens when the learned function is too specific to the training set. This problem is called Overfitting. And usually appears when the data is too heterogeneous and the dimensionality is too high. Figure 1.3 illustrates the learning process.

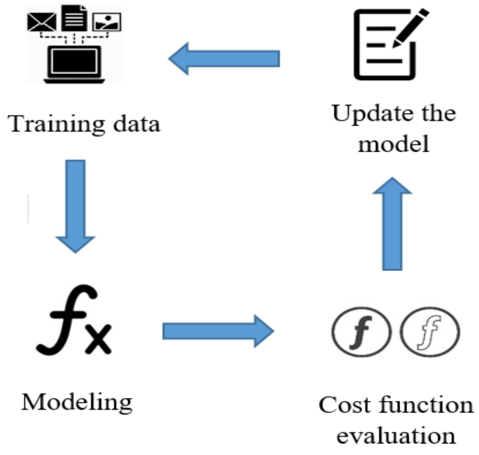


Figure 1.3- Model learning steps

1.2.3 Suport Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised machine learning algorithm that can be used for both classification and regression analysis [26]. Nevertheless, it is often used for classification objectives. SVM builds a model that can segregate with an extreme margin a n-dimensional space into classes. Consider Figure 1.4 diagram which we have two different classes that are separated using a decision boundary (also called hyperplane)

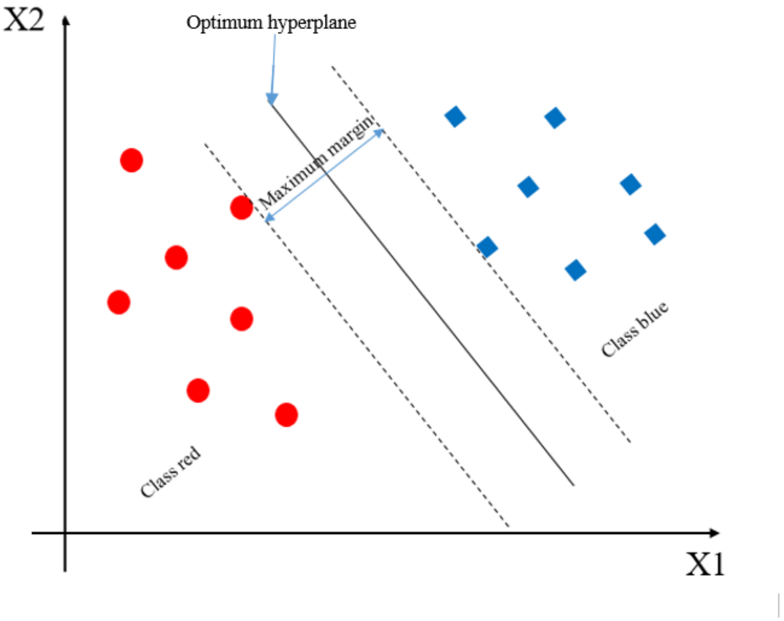


Figure 1.4- Showcase of SVM hyperplane in case of 2-dimension problem

There are two types of SVM:

-Linear SVM: Linear SVM is used when we can separate different classes of data using a single straight line (in the case of 2D space). Take for example figure 1.4. In order to facilitate. Assuming that the input data contains only two features X1 and X2. Hence, the Linear SVM helps to find the best decision boundary to classify the pairs (X1, X2) in either red or blue classes.

Non-linear SVM: Non-linear SVM is used when the data classes cannot be separated using a straight line (see Figure 1.5). The way for SVM to segregate non-linear data is by applying the Kernel trick [27]. So there is a function that transforms for all X_i of input data in space Ӽ an inner product in another space Ѵ for mapping the inputs into n-dimensional feature space.

We have φ: Ӽ $\longrightarrow$ Ѵ

Which satisfies: $K(X_i, X_j) = <\varphi(X_i), \varphi\ (X_j)>$ Ѵ

However, working in high-dimensional feature space increases the generalization error of the SVM model. Hence, much more training data is required to perform well [28]. Nevertheless, the SVM algorithm has been widely used in material science. Notably applied to assist the design and fabrication of solar cells [29, 30].

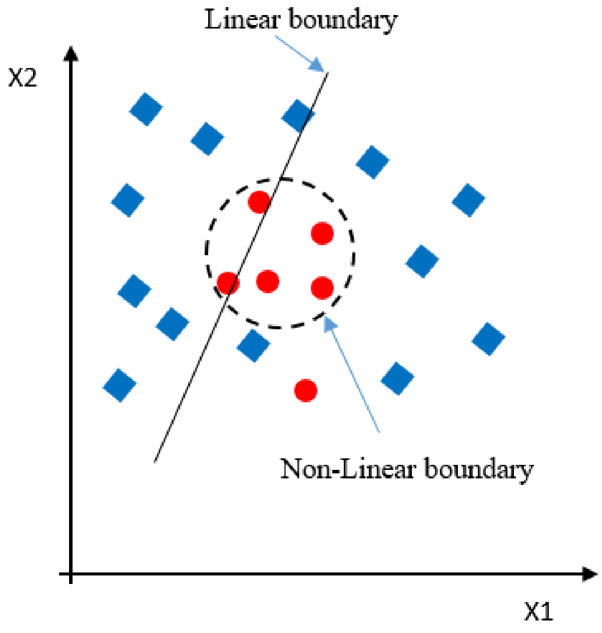


Figure 1.5- Example of nonlinear separable data.

From Figure 1.5 we can observe that the Linear boundary cannot separate the blue class from the red class. Hence, the probability of getting an incorrect prediction increases. While the Non-Linear boundary facilitates the separation between the two classes. Therefore, it is critical to know the type of your data, if is it Linear or Non-Linear separable to correctly choose the appropriate model. In the next chapter, we will see a practical application of this problem.

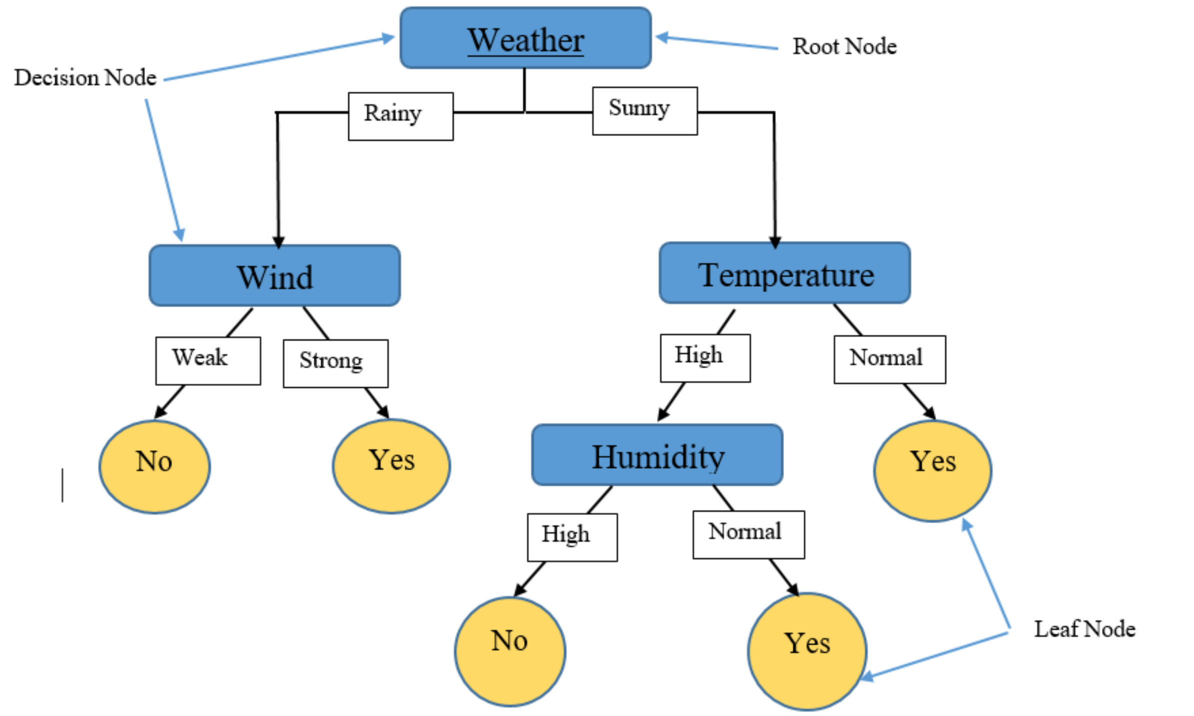


Figure 1.6- Diagram explains the general structure of a decision tree.

1.2.4 Decision Tree (DT)

Decision Tree development began in 1959 in a paper titled "*Matching and Prediction on the Principle of Biological Classification*" [31]. In late 1970s. Breiman et all invented the CART version (a space of regression tree), which would become later a world standard for decision tree analysis [32]. DT is considered a supervised learning method used for both classification and regression problems. As it relies on a tree structure with two types of nodes, a Decision Node and a Leaf Node (see Figure 1.6). The leaf node represents the outcome of the model.

In particular, given an example of labeled data (X_n, Y_n), with n ∈ {1, …., N}. The Decision Tree algorithm creates a non-linear decision separator by using several linear separators. As in SVM, these boundaries take the form of a hyperplane which can be written as: X_l = c. where I represents input features. l ∈ {1, …., L},

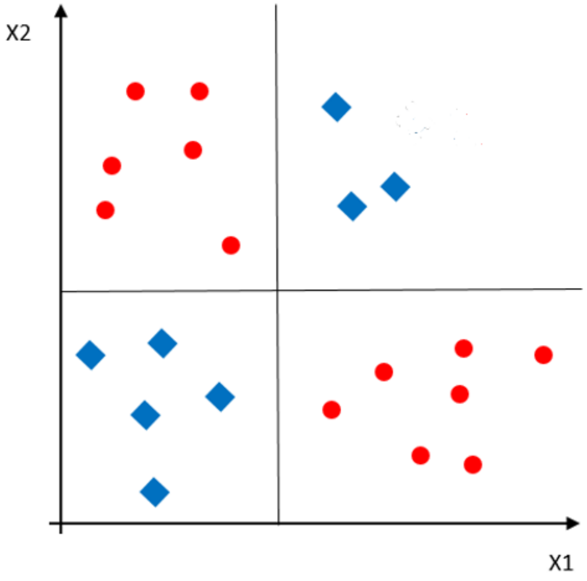


Figure 1.7- Example of a decision boundary made by a Decision Tree. Each hyperplane represents a decision node

Moreover, Figure 1.7 shows an example of a hyperplane separator in the case of 2D input data (two features: X_1, X_2), with two different classes represented by red and blue colors (two labels: Y_1, Y_2). This example is a task for a classification problem, in the case of regression problem $Y_n \in \mathbb{R}$. Each node of the decision tree is associated with a hyperplane.

The following terminologies in Figure 1.6 help to explain this concept better:

Root Node: The root node is the start of the decision tree that represents the entire dataset. Which will be divided into two or more sets.

Splitting: Splitting is the dividing process of the decision nodes into child nodes by using splitting criteria.

Parent/child node: The root node represents the parent node. Where the other nodes are called child nodes.

Leaf node: It represents the last node which the tree cannot divide further. It also gives the final output of the model.

Usually. A Decision Tree performs poorly in its basic form in complex problems. Therefore, in this research, we used the ensemble method based on the Decision Tree technique which often can achieve better results compared to a single decision tree approach. In particular, the ensemble learning algorithms used in this research are Extra Trees (ET), eXtreme Gradient Boost (XGBoost), and Random Forest (RF). We will discuss the differences between these algorithms and their use case further in the following chapters. However, each ML technique has its costs and benefits depending on the problem assigned, and there are no algorithms that are valid for every problem. Therefore, our approach relies on trying different ML techniques and then choosing the most convenient technique for a particular problem.

1.2.5 Concept of Artificial Neural Networks (ANN) and Deep Learning

Table 1.1- Analogy between biological and artificial neural networks

BIOLOGICAL NEURAL NETWORK	ARTIFICIAL NEURAL NETWORK
NEURON	Node
DENDRITE	Input
AXON	Output
CONNECTION OF NEURONS	Connection Weight

Artificial Neural Networks (ANN) were introduced as an implementation form of ML that can imitate the way the human brain works [25]. This model is based on a collection of nodes called Artificial Neurons which mimics the mechanism of biological neurons. Table 1.1 illustrates the analogy between the artificial NN and the brain. The output of each artificial neuron is calculated using a non-linear function called the Activation function. There are several types of activation functions depending on the task of the neuron. The following example shown in Figure 1.8 illustrates an ANN with a single node and three inputs.

The input signal {x_i} is multiplied by a coefficients called weights {w} associated with another factor b called bias producing what is known as weighed sum (in this step the concept is similar to the supervised ML) which is calculated as follows:

$$v = \sum_i^n w_i \, x_i + b \qquad (3)$$

Then, the weighted sum enters the node and yields the output by using the activation function as follows:

$$y = f(v)$$

A single node by itself usually doesn't perform well. However, their interconnection allows to extraction of higher-level patterns from the data. Therefore, Deep learning was introduced, which represents the use of multiple nodes in addition to multiple layers called hidden layers, as shown in Figure 1.9. Furthermore, there are two main types of neural network models depending on the learning method:

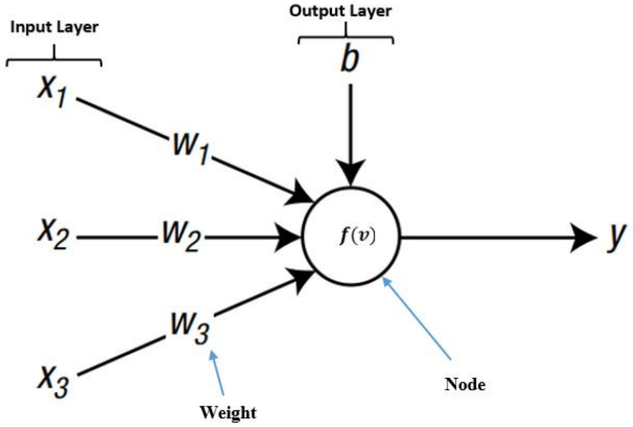


Figure 1.8- Example of a simple artificial neuron (Perceptron)

A Feedforward network (FNN): is considered the simplest implementation of NN. In this method, the signal (information) moves from the input layer to the output layer. There is no back forward to the previous nodes [33].

Recurrent network (RNN): this method creates a cycle allowing some nodes to update their attached inputs. This feedback allows to reuse of the output data to affect the earlier learning stages [34].

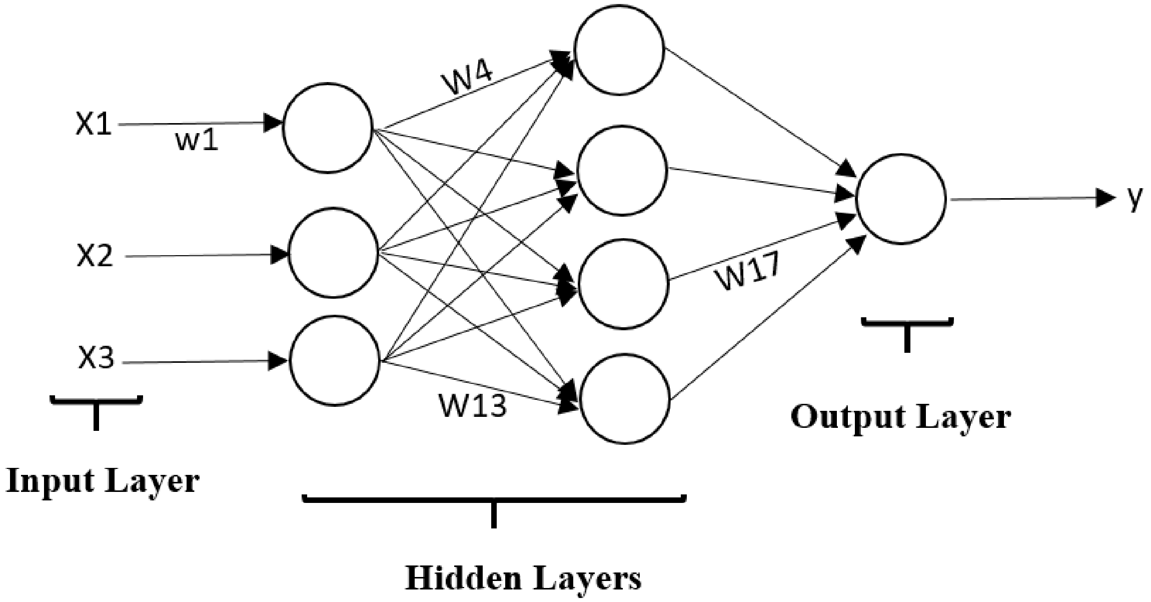


Figure 1.9- Example of Deep learning neural network

You can easily find some examples using the concepts of ML and Deep Learning interchangeably. In general, the relation between ML, Deep Learning, and AI is as follows: "Deep Learning is a kind of Machine learning, and Machine learning is a kind of Artificial Intelligence” [25].

1.2.6 Definition of some main concepts

Overfitting: this phenomenon is considered the primary cause of the model failure of the generalization process for new data. This phenomenon occurs when the model learns the detail and noise from the training data (fit the training data perfectly), and prevents the model from reflecting correctly the general behavior. Figure 1.10 gives a comparison between overfitting and underfitting.

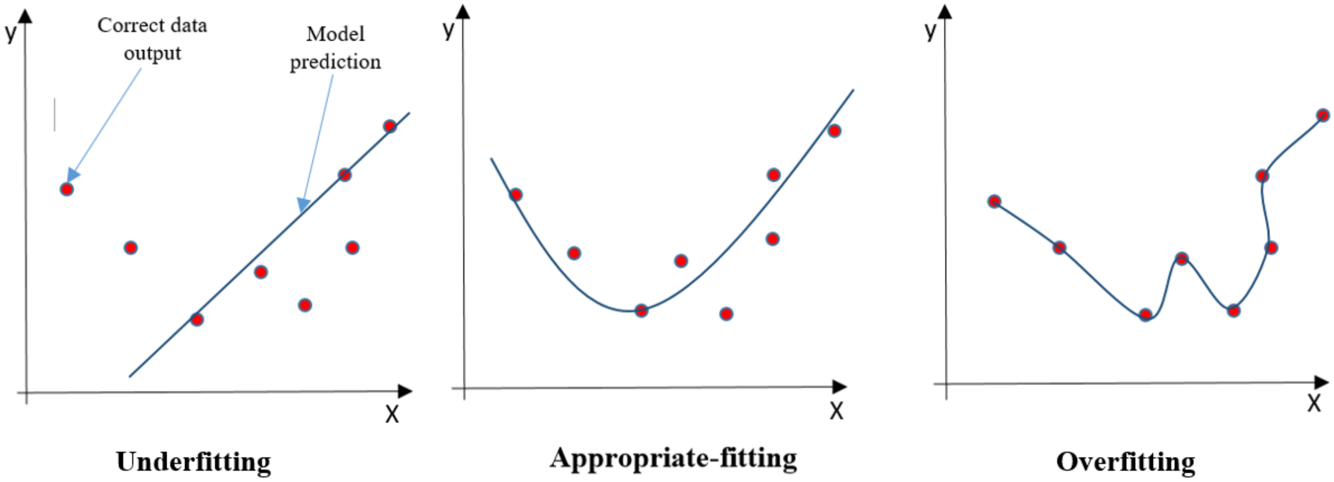


Figure 1.10- Overfitting and underfitting in machine learning

From Figure 1.10, we can see that the Overfitted model cannot predict correctly new data because that model seems unsystematic and failed to extract a particular pattern from the training data. Where the Underfitted model failed to fit both training and new data. Obviously, the appropriate model is between underfitting and overfitting training, even though that model is not 100% correct, but it gives a very close prediction of the true values.

Noise in data: This term represents the data samples that have meaningless information in the machine learning process or include corrupted data. Noisy data can affect the learning process and confuse the ML model.

Bias: Is a phenomenon in the ML model that yield the results to diverge from the correct value due to some features in the dataset being given incorrect weight. This occurs due to the false assumptions during the machine learning process.

1.3 Applications of ML in Material Science and Related works

In this section, we outlined several works that have applied machine learning techniques in the material science field. There are various applications of machine learning depending on the purpose of the study. Usually, the objective can be summed up as follows:

Material discovery: Under this category, there are properties prediction, active learning, and inverse design [35] (this research falls into this category).

Properties prediction models aim to estimate the uncovered material properties based on their structure or chemical composition. We will see an application of that in chapter 5, where we predict the perovskite band gap using the structure and the components of the perovskite solar cell. Moreover, regarding new materials, this technique usually can run orders faster than ab initio simulations or experiments [35]. In fact, this will significantly accelerate material discovery. Moreover, it has been applied successfully to accelerate the development of organic light-emitting diodes [36], and lithium-ion batteries [37]. These review papers discussed several applications [29, 38, 39].

The objective of the active learning models is to discover the complex material space. Such as the configuration of atomic structure using Bayesian active learning [40]. Whereas, inverse design aims to predict the material structure that assists in optimizing the performance of a device based on existing data. For instance, in Chapter 4 we estimated the optimum structure for achieving high stability in PSV devices.

Interpretation and Visualization: The goal of interpretation and visualization models is to help understand complex material spaces and discover the main factors that can potentially assist in finding new theories for materials design. For example, mapping the uncharted territory in ice structures [41].

Chapter 2. A comparison of different Machine Learning and artificial neural networks techniques for forecasting perovskite solar cells (PSCs) stability

2.1 Introduction

In this chapter, we present a comparison between three machine learning and neural networks (NN) techniques depending on their performance in forecasting perovskite solar cell stability. The ML and NN models are trained using experimental data from 377 solar cell samples collected from previous works. The three algorithms used in this chapter are: SVM, Multilayer perceptron (MLP), and Probabilistic Neural Networks (PNN). However, SVM and MLP techniques are widely used in material science and photovoltaic research. For instance, M.Pan et al used SVM for predicting the photovoltaic power for an ultra-short term [42], while MLP is commonly used in fault identification. For example, F. Khondoker et al used MLP for photovoltaic panel array fault prediction [43]. Whereas, the PNN is much less used due to the property of this algorithm which is hard to implement. Case in point, R.G.Vieira et al published a work involving a comparison between multilayer perceptron and probabilistic neural networks for photovoltaic system fault detection [44].

In particular, we will first provide an identification of the perovskite solar cell and the ML techniques used in this chapter. Dataset collection and organization also will be clarified. Then, we will discuss the model building and the hyperparameter tuning. Further, we will explore the dataset, and we will discuss and analyze the model's obtained results. Moreover, we compared these results with experimental data to investigate the most suitable algorithm for the intended goal. Finally, we will close the chapter with our proposition for achieving better results and give a brief outlook of this approach.

2.2 Data collection and preparation

2.2.1 Perovskite solar cell description

The perovskite solar cell is considered a third-generation solar cell based on emerging thin film technology [45]. It was discovered by Tsutomu Miyasaka in 2009 by modifying the dye-sensitized solar cell device structure with a perovskite thin layer deposited onto titanium oxide (TiO_2) [46, 47]. The power conversion efficiency generated was only 3.8%. However, the stability of the cell performs very poorly due to the presence of a liquid electrolyte. In 2012, the first solid-state PSC was fabricated using $MAPbI_3$ deposited onto a thick mesoporous TiO_2 layer with Spero-MeOTAD as a hole-transport layer and Au as back contact [48]. Figure 2.1 represents the development of the PSC PCE through the past decade, this figure shows only the highest PCE recorded in each year taken from the National Renewable Energy Laboratory (NREL) [16].

The structure of the PSCs is classified into two types depending on the nature of the transport material. If the electron transport layer (ETL) encounters the light rays first, then the PSCs are classified as "Regular n-i-p structures" (or Normal structures). If the hole transport layer (HTL) encounters the light first, then it's called an "Inverted p-i-n structure". The basic regular PSC consists of an ETL layer deposited on glass (e.g. FTO: fluorine-doped tinoxide). Then an absorber layer has a Perovskite structure called an active layer. An HTL and Back Contact. Figure 2.2 presents the components of a regular perovskite solar cell.

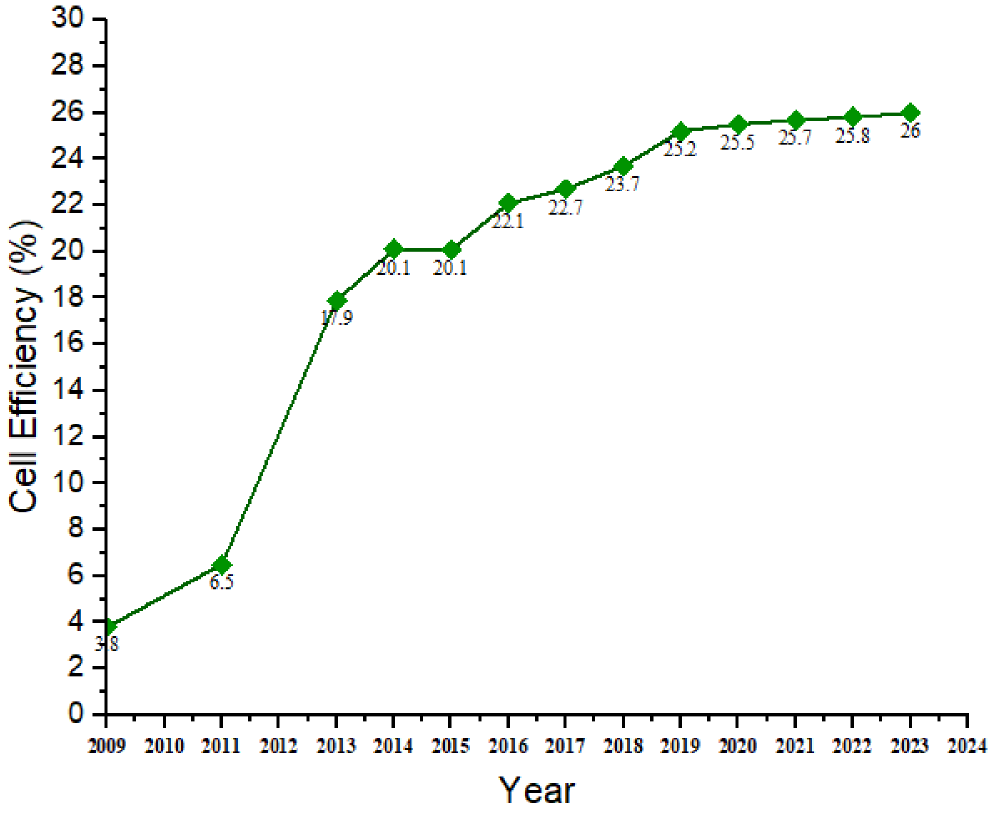


Figure 2.1- Highest PCE recorded of PSV every year.

The active layer (absorbent) of the PSC consists of a perovskite structure materials with ABX_3 form. The A site represents an organic cation, whereas the B site consists of a metallic cation, the X site is a halogen anion such as iodide, chloride, or bromide. However, even though the organic materials are cheap and easily tunable, although they still struggle with chemical instability against environmental factors like moisture, which causes the stability degradation of the device. Therefore, many researchers are shifting into using all-inorganic perovskites by replacing the organic cation with inorganic materials like Cesium (Cs) due to their potential to be more resistant to environmental factors compared to the hybrid organic–inorganic counterparts [49]. Moreover, the perovskite materials can be comprised of multi-cation and/or multi-anion compositions.

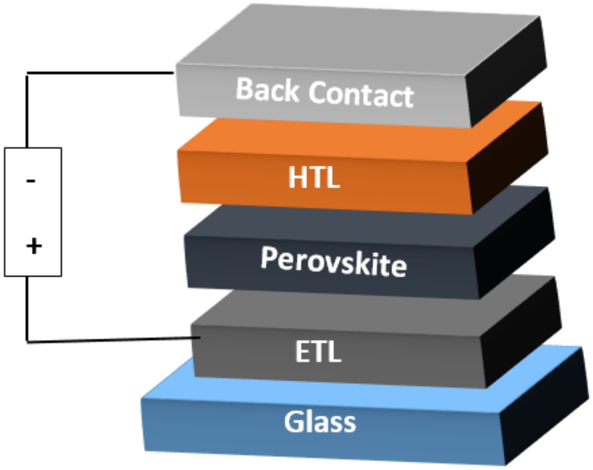


Figure 2.2- Regular n-i-p perovskite structure

The ETL layer improves the efficiency of extracting the generated electrons by the perovskite active layer. Also, it reduces the recombination of the charges by blocking the hole from migrating to the counter electrode. Moreover, the electron transport materials (ETM) can be organic or inorganic. For instance, TiO_2 is an inorganic material that is widely used as ETL, and PCBM (one of the fullerene derivatives, -phenyl-C61-butyric acid methyl ester) is an organic material found that it has a better ability of charge extraction than TiO_2 in inverted structure PSC [50].

In contrast, the HTL collects and transports the hole from the perovskite active layer to the electrode to increase the electron-hole separation. Moreover, the highest occupied molecular orbit (HOMO) in the hole transport materials (HTMs) should correspond to the valence band of the perovskite material to allow hole transport [50]. Similar to the ETM, the HTM is divided into two categories: organic, like the widely used spiro-OMeTAD, and inorganic like NiO. Furthermore, each layer of material can be doped with organic or inorganic materials to change its proprieties for different purposes.

However, because the PSC has been developed in the past decade where a huge amount of data is being shared on the internet, the effort and time-consuming process of collecting labeled data with experiments has been minimized. Nowadays, many different laboratories and research teams share their experiment results on different online platforms. Therefore. The fact that the PSC devices are progressively developing and the availability of labeled data makes the PSC the most appropriate device for ML applications in this research.

2.2.2 Construction of dataset

In this chapter, a large dataset containing 377 experimental data samples of regular PSC devices was used to train and test different ML and NN models. The data was collected through screening several papers that provide the measurement of the PSCs stability degradation under various environmental conditions. The majority of this data was obtained from the Perovskite Database Project Team [51]. The dataset consists of the materials composed of the PSC device layers shown in Figure 2.2. In addition to the temperature, light intensity, relative humidity, and atmosphere conditions of the environment where the PSC devices are stored, these features represent the inputs for the ML models. Furthermore, the device stability T80 is considered as the output for the models. i.e. the value of the stability measured for which the device PCE degrades down to 80% of its initial value.

In particular, the relative humidity values have been classified into three ranges: from 0% to 30% RH, 31% to 55% RH, and above 55% RH. The details for this classification are provided in the next chapter (section 3.3). Moreover, we will see further that the encapsulation of the solar cell may slightly enhance the device's stability. Therefore, because the majority of the PSC samples are not encapsulated, we have excluded the encapsulated cells from this analysis to prevent this additional protection (encapsulation) against environmental conditions that may affect the ML model. Table 2.1 represents a sample of the dataset used to train the ML models.

Table 2.1- Perovskite solar cells stability measurements dataset

Features *Cells*	*Glass*	*ETL*	*Perovskite*	*....*	*Back Contact*	*Target*	*Stability (days)*
Cell 1	SLG | FTO	PCBM-60 | BCP	MAPbI	….	Ag		1
Cell 2	SLG | ITO	TiO2-c	FAMAPbBrI	….	Au		18
Cell 3	SLG | FTO	PCBM-60	MAPbI	….	Al		13
Cell 4	SLG | FTO	TiO2-c | TiO2-mp	CsPbBrI	….	Ag | MoO3		55
Cell 5	SLG | ITO	SnO2-nt	FAPbBrI	….	Cu		15

2.3 Materials and Methods

2.3.1 Multilayer Perceptron and Probabilistic Neural Networks

Multilayer perceptron (MLP) is an implementation of a feedforward type of artificial neural network that has at least three layers of nodes: an input layer, a hidden layer, and an output layer, and each node has a nonlinear activation function [24]. Hence, it has a significant advantage for problems consisting of nonlinearly separable data. The MLP algorithm can be used for both regression and classification tasks.

Probabilistic Neural Networks (PNN) (also known as Bayesian networks) is a feedforward type of artificial neural network derived from the Bayes decision strategy, which utilizes the sum of identical isotropy Gaussians to determine the likelihood function of a given class [52]. In particular, the PNN algorithm is used for classification problems which is considered as an excellent pattern classifier. Moreover, the PNN uses prior hypotheses to improve the output predictions and provides each class with a probability density function (PDF) for each input being an element of that class [53]. By using the Parzan estimator, for a class (i) the PDF can be estimated as follows:

Likelihood function for class i is: $L_i(x) = \left(\frac{1}{N_i}\right)\left(\frac{1}{(2\pi\sigma)^{\frac{k}{2}}}\right)e^{-(x-x_i)^2/\sigma}$ (4)

N is the number of the training samples in class i, and k is the dimension of input nodes. σ is the variance of the Gaussians that must optimized during the training process (similar to w and b from ordinary ANN).

And the conditional probability for class i is: $P_i(x) = \frac{L_i(x)}{\sum_i^M L_i(x)}$ (5)

Where M is the number of classes. The training process for this type of NN is fairly rapid but can require a large set of data. Figure 2.3 shows an example of PNN networks with two classes.

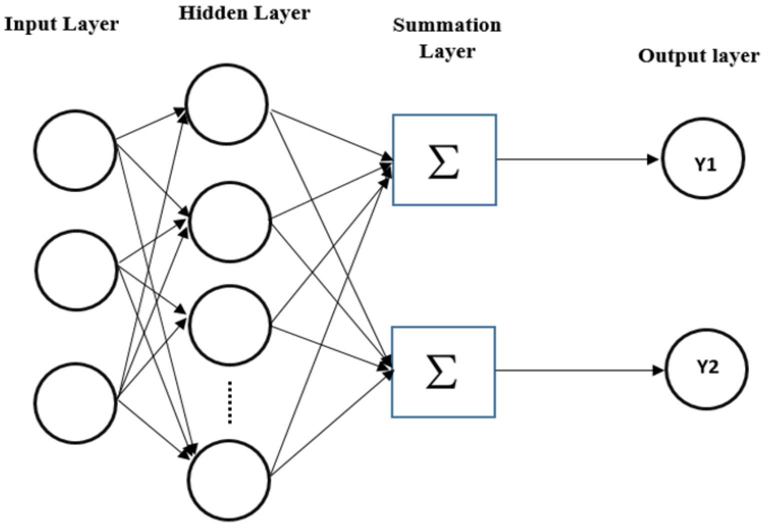


Figure 2.3- Network of PNN.

2.3.2 Data pre-processing

Data pre-processing reflects the process of the manipulation and preparation performed on raw data to be tailored in a format appropriate for machine learning applications. It is used as well to enhance the ML model performance. However, there are several different techniques used for preprocessing data. For instance, data cleaning, transformation, imputation, feature scaling (standardization, normalization, etc.), feature selection, etc.

Foremost, we have performed a manual exploration of the dataset to uncover the initial patterns, check missing values, and correct typing mistakes. Also, we have organized the data in the form of a table (matrix), in which the rows represent a sample of solar cells and the columns represent the data features (e.g. ELT layer, temperature ...etc.) as shown in table 2.1. Practically, in all cases, the material dataset contains missing data that are not provided from the original work. For example, some of the papers from which the data was collected did not mention the environmental temperature or humidity. This missing data in the dataset often creates a huge problem for the ML model. Hence, it is crucial to handle the missing values. Generally, there are two ways to solve this problem:

-Deleting the particular sample: This method is used commonly to handle the null values, or sometimes it is used when one sample contains several missing values.

-Calculating the mean: This method is based on calculating the mean of the column where the missing value is, then replacing the missing value with the mean value of that column. For example, a sample has an unknown temperature. In this way, we calculate the mean value of all the temperatures existing in the dataset and then replace the missing value with the mean value. This method is widely used and very useful for features that have numerical values.

However, we adopted a different approach for preserving the maximum of the data. Based on the fact that this dataset consists of experimental data conducted within laboratories. Hence, we assumed that the missing data for the stability degradation measurements was taken under ambient conditions. Moreover, usually, if there are any changes from ambient conditions it will be notified in the published paper of the concerned work. Therefore, we have compensated the temperature and humidity values with 25 ^{o}C for the temperature and 0%-30% RH for the humidity.

Furthermore, the ML algorithms work completely with mathematics and receive only numerical values as input. Hence, we have encoded the features containing characters by creating a sparse matrix of binary columns, and each category (element) is represented by a binary value. i.e. replaced by 1 if that category exists and 0 if not. This process was done using OneHotEncoder (OHE) from the Scikit Learn library [38]. The reason for choosing this approach instead of other approaches (e.g. Label Encoder which replaces a category by a random number) is the attempt to neutralize the ordinal relationship between the substitute variables. However, after the encoding step, the dataset consists of binary data that represents the materials components and other numerical data like temperature and light intensity which scale much bigger than the binary values. This margin provokes the bias phenomena which means that the features with large values dominate the model. Therefore, to ensure that all the features contribute correctly we used standardization via scaling to unite variance in order to put the variables in the same range. The standard score of sample i is calculated as follows:

$$i' = \frac{(i-u)}{s} \qquad (6)$$

Where i$^{'}$ is the standardized value of i, u represents the mean of the training samples, and s is the standard deviation of the dataset. By applying this method the data can be transformed to a more consistent scale and prevent the bias phenomena from occurring, also it makes the regression models learn the patterns from the data easier. Figure 2.4 illustrates both the encoding and standardization processes.

Finally, the dataset was divided into two subsets. 85% of the data was directed to train the ML models, and 15% for testing the models. Note that each one of the training dataset and the test dataset consists of an input and output pair {x, y}.

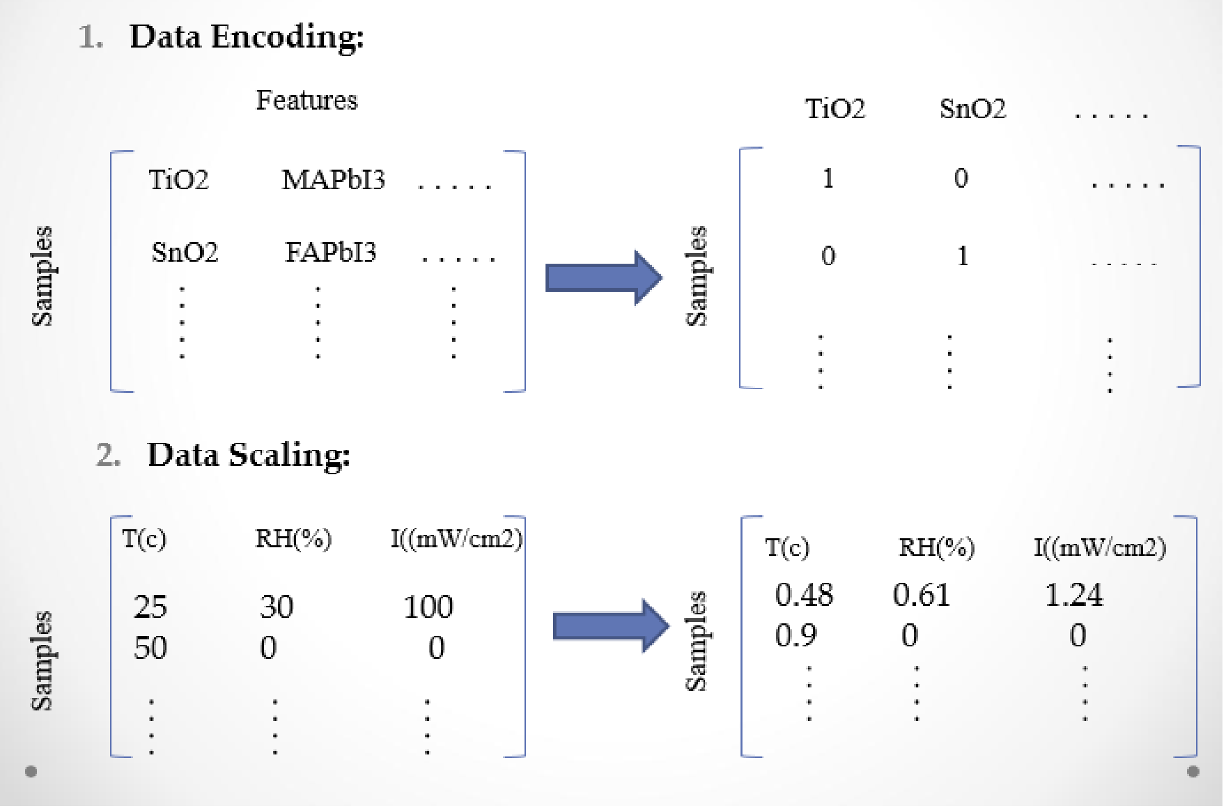


Figure 2.4- Data encoding and standardization processes.

2.3.3 Model configurations

This chapter aims to perform multiple machine learning experiments to identify the optimum model for our material dataset. However, this process requires choosing the optimal parameters that make the learning algorithm correctly map the input features to the output. Generally, the parameters that control the learning process and the resulting model parameters are called Hyperparameters. For instance, some common hyperparameters we have: optimization algorithm, the activation function in case of NN, the choice of the loss and cost functions, etc.

However, the common method for choosing the optimum hyperparameters is by modifying the values of different parameters and repeating the experiment until the best results are obtained. Therefore, we used a technique called GridSearchCV from Scikit Learn for this purpose. In particular, GridSearchCV performs hyperparameter tuning to determine the optimal value by processing a given set of parameters in the form of a grid. Moreover, GridSearchCV applies every combination possible of the parameters in the grid, then it uses an internal cross-validation technique to calculate the score for each combination. The bigger the score obtained the better that combination of hyperparameters. Figure 2.5 illustrates the steps of hyperparameter tuning used by GridSearchCV.

Depending on the previous approach, the optimal hyperparameters of each of the SVM and MLPRegressor algorithms are found as follows:

-SVM: -The kernel type: Radial Basis Function

-The degree of the polynomial kernel functions = 1

-MLP regressor: -The number of hidden layer = 5

-The number of neurons in the ith hidden layer = 200

- The number of iterations of the training process = 200

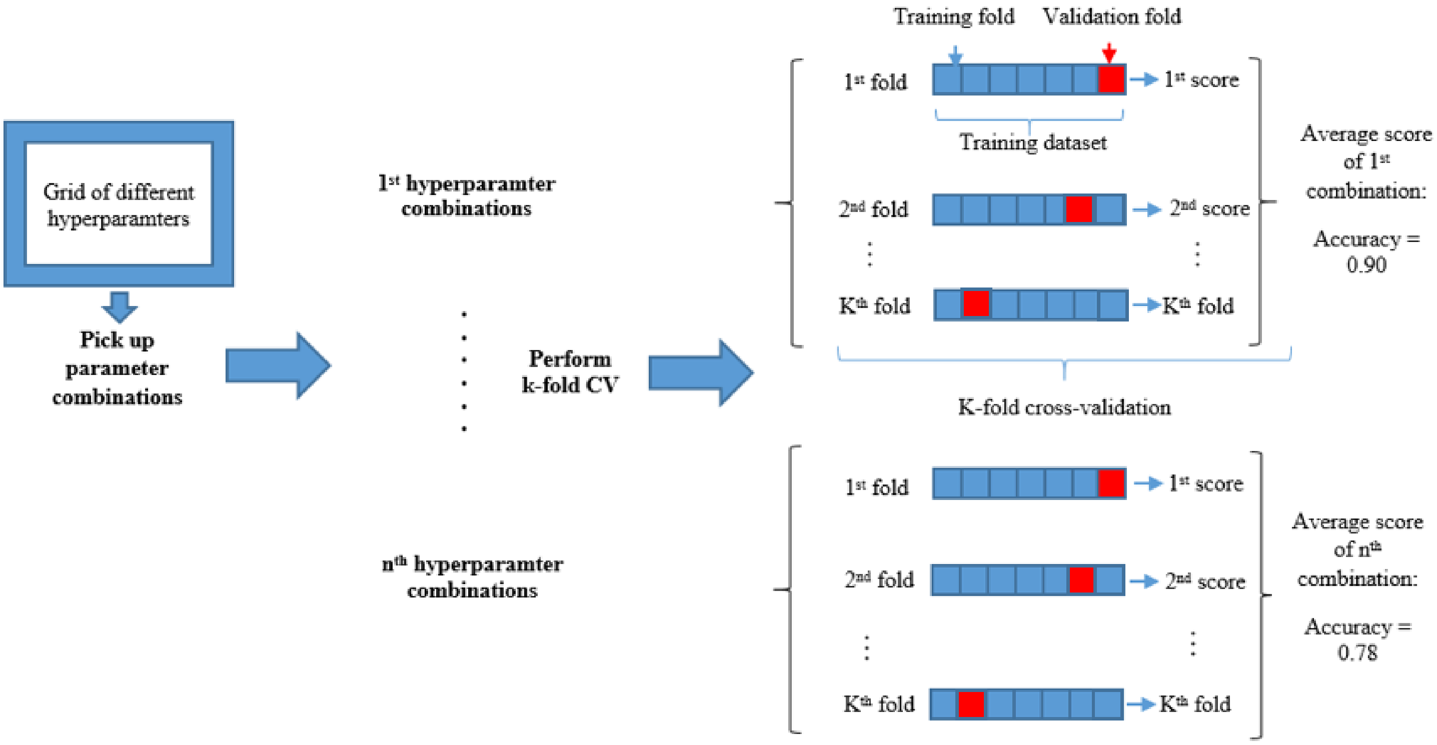


Figure 2.5- Hyperparameter tuning by GridSearchCV

However, in the case of the PNN algorithm, a different approach was implemented. Given the specific structure of this algorithm, we have used a library from tensorflow [55] for the probabilistic calculations to make the prior and posterior functions. Then, it is employed in the standard algorithm of Neural Networks. The hyperparameters of the NN are as follows:

-The loss function is Mean Square Error (MSE): $MSE = \frac{\sum(y_i - prediction_i)}{n}$ (7)

Where y is the correct output and n is the number of total outputs. The Root Mean Squared Propagation (RMSprop) was used as an optimizer, and the learning rate = 0.0001.

Furthermore, to develop the proposed experiment, the life cycle of the models can be summarized as follows: (i) tuning the model hyperparameters using the GridSearchCV algorithm,

(ii) compiling and training the model, (iii) using the test set to assess the model accuracy, (iv) use the model to make new predictions. However, we used the same training set and test set to train and evaluate all the models to ensure fair comparison analyses.

2.4 Results and Discussion

The test dataset contains 55 experimental data samples. However, to assess the model's performance concerning our PSC dataset, we have plotted the stability at T80 from the test dataset along with each model prediction. Knowing that the same inputs from the test dataset were used to make the predictions for all the models. The resulting curves are illustrated in Figures 2.6, 2.7, and 3.8. where the blue curve represents the real values and the red curve represents the model's prediction results.

Table 2.2 contains the accuracy score of each algorithm. The accuracy was calculated by using: the training set and then by using the test set. Moreover, the Metric and Score library from Sklearn was used for this calculation, this algorithm is based on a cross-validation algorithm that computes the accuracy score by using the fraction of correct predictions method which can be expressed by the following function:

$$accuracy(y, pred) = \frac{1}{n} \sum_{i=0}^{n-1} 1(pred_i = y_i) \quad (8)$$

Where $pred_i$ is the predicted value of the i-th sample and y_i is the corresponding real value, n is the number of samples.

Table 2.2- Stability prediction accuracy of PSCs from different ML models.

Algorithms	**Test dataset prediction accuracy**	**Training dataset prediction accuracy**
SVM regressor	17%	24.78%
MLP regressor	63.1%	71%
PNN	70.17%	81.24%

As we can note from Figure 2.6, the SVM algorithm precision was very low. From Table 2.2, this model perform only 24.78% in the case of training data prediction, and only 17% of the test data was predicted correctly. However, the SVM model in this case is considered a very simple model, which is based on a linear polynomial function to fit the input with the output. Regardless that this function was found the best method by using GridSearchCV, the SVM technique can apply the kernel trick to transform the data from the original space to another feature space which can help the algorithm to solve more complex problems [57]. However, as we have seen in the previous chapter (section 1.2.3-support vector machine), the kernel trick requires much more training data to produce better results. In particular, the main reason for this poor performance of the SVM algorithm is due to the quality of the dataset. The datasets based upon the material data are generally considered very heterogeneous and nonlinear separable. Moreover, the SVM algorithms will underperform in the case when the number of descriptors (features) is around or higher than the number of data samples [57]. In our case, in the preprocessing stage, the OHE transforms the 387 material types from the dataset to features. Hence, the results of this operation produced 391 features (the material types + the environmental conditions) against 377 data samples which guarantees the SVM model's underperformance. Furthermore, during the manual exploration of the data, we noticed that there is some noise in the data, which some PSC devices contain the same structure and environmental conditions but different stability. Therefore, the SVM algorithm doesn't perform well in case the dataset contains much noise [57].

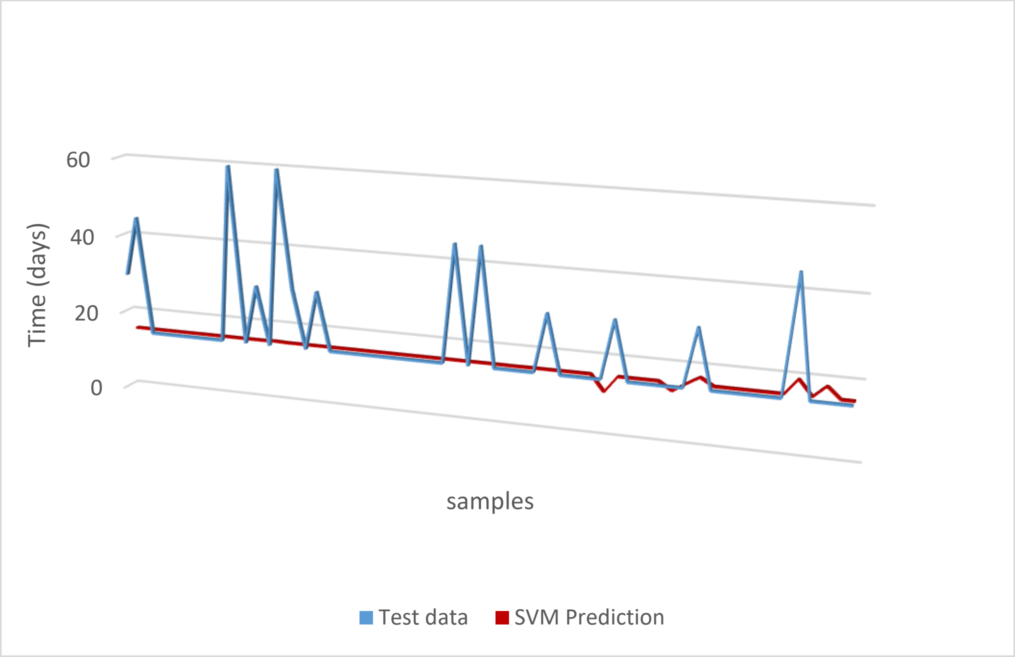


Figure 2.6- Comparison between the stability predicted using the SVM model against experimental values.

Figure 2.7 shows that the multilayer perceptron regressor performs much better than SVM. From Table 2.2, the MLPRegressor accuracy is 71% for the training data and 63.1% for the test data. Hence, the growth of the accuracy was 46% for both test and training sets. However, in the case of the Neural Networks algorithms, the hyperparameters of these kind of models is quite hard to optimize [56]. Even though the dataset contains noises, the MLPRegressor gives relatively good results. However, this performance enhancement may due to the fact that the Neural network models are usually addressed for nonlinearly separable data [58] which is the case of our problem.

In particular, the reason behind the gap between the real values and the predicted values may be due to that the input features used in this work do not represent adequately the relationship between the input and the outputs. Hence, the MLP model cannot extract sufficient information from the training data. Moreover, the size of the dataset and the noise in the data also cause a degradation of the model performance. In general, the NN models are considered as a black box and it is not human interpretable. Simply, it is very hard to know why or how this model produces a specific output.

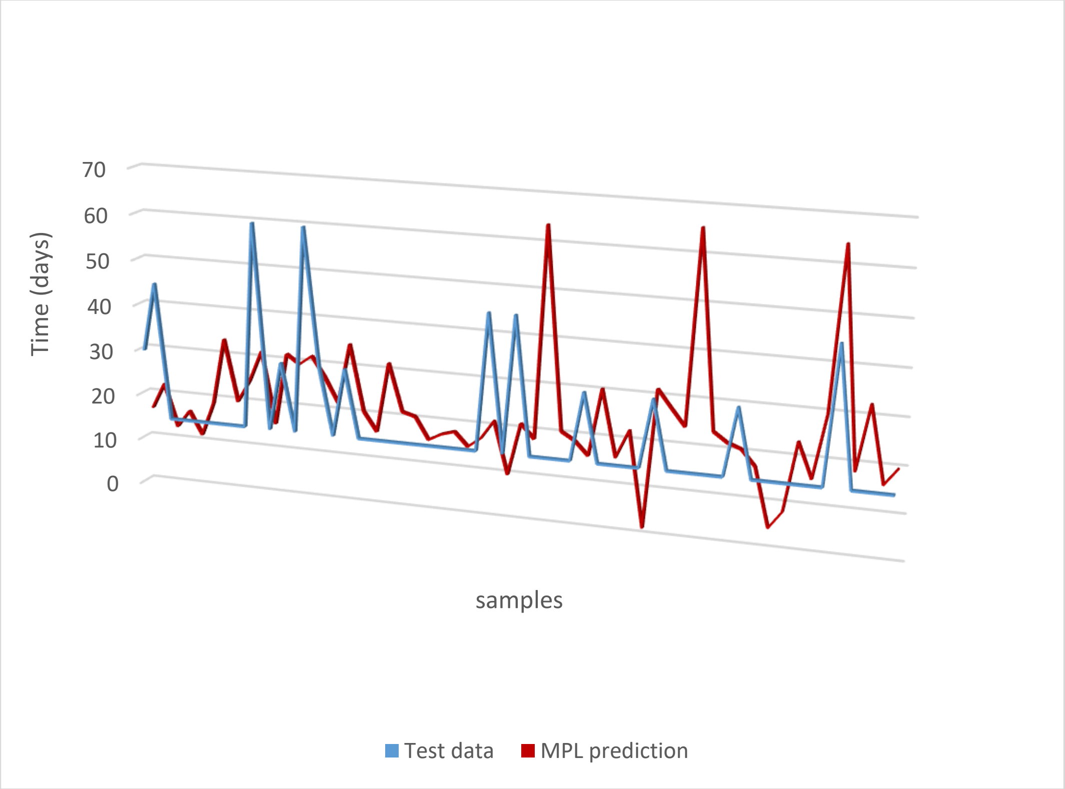


Figure 2.7- Comparison between the stability predicted using MLP regressor against experimental values

Figure 2.8 shows that the curve of the stability predicted value is nearly identical to the curve of the real values. Which makes the PNN behalf as the best model compared to the two previous models. The probabilistic approach added to the neural networks can empower the model to handle much better the uncertainty in the dataset caused by insufficient information and the noise in the data [40]. Moreover, this model can learn better from small-size datasets compared to artificial neural networks. Table 2.2 illustrates that the PNN model gives the highest accuracy both for the training set and test set, which scores 81.24% in the training data and 70.17% in the test data. However, the results of this model are also considered probabilistic, which means that the same inputs may produce a slightly different output. In fact, this feature can be considered as an advantage and disadvantage at the same time. Obviously, a model that sometimes gives different outputs from the same inputs is considered a bad model. However, from an analytic perspective, this behavior suggests that the output is unlikely correct. So it is better to get "I don't know" from the model rather than getting a misleading prediction.

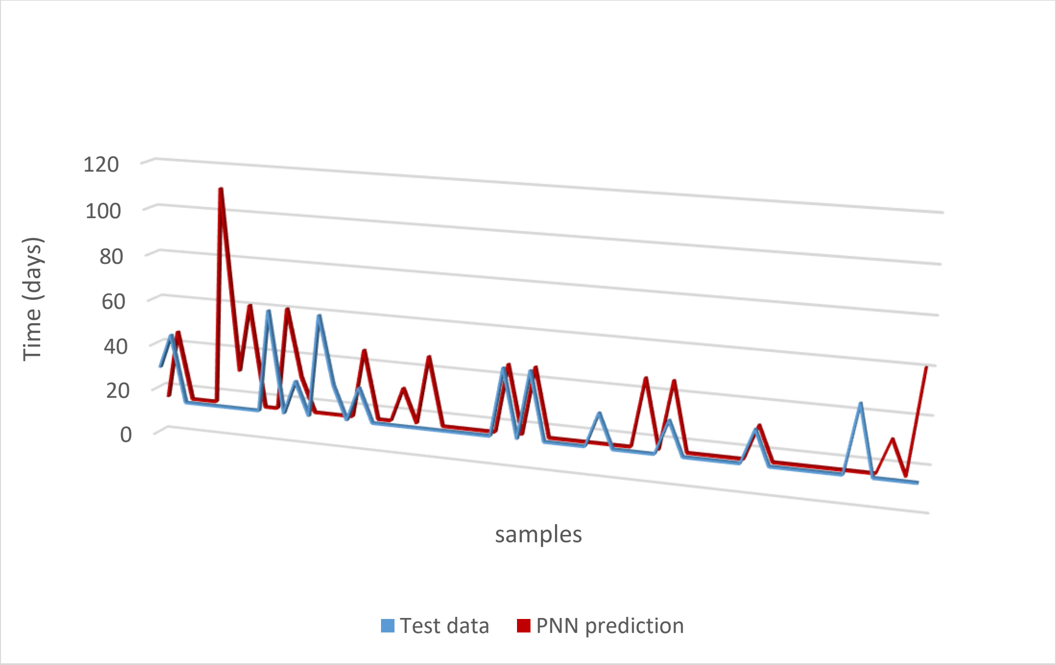


Figure 2.8- Comparison between the stability predicted using PNN against experimental values

2.5 Chapter summary and outlook

In this chapter, we have identified three ML/NN algorithms: SVM, MLP, and PNN, which were used for the prediction of the PSC device stability. The objective of this experiment was to compare various ML/NN techniques in order to pick up the most appropriate one for our problem. In particular, the advantages and disadvantages of these techniques can be summarized in Table 2.3. In conclusion, based on the results obtained, the neural network algorithms are considered advisable for the assigned task. The probabilistic neural networks give higher accuracy compared to the MLP neural network algorithm. However, the PNN has a critical limitation in usage. It is used only for classification tasks. This obliged us to classify the dataset into many classes so that we could compare it with the other techniques. Although this method is not practical for regression tasks. In this context, the SVM shows a poor performance when it comes to heterogeneous data. However, the observation of this experiment gives some conclusions:

It was found that the methods of the preprocessing stage are specific for each algorithm. This means that if a specific method worked for one algorithm, it cannot be generalized for all the algorithms. For example, encoding the categorical data with OHE usually gives good results in the case of NN algorithms, but we have seen that it causes an underperformance for the SVM algorithm, notably in the case when the number of categories is large.

Furthermore, the uncertainty and noise in the dataset can underperform the ML/NN models. Therefore, the problem of handling noise in prediction applications needs to be investigated further.

Table 2.3- Advantages and disadvantages of the ML/NN algorithms used in this chapter.

Algorithms	Advantages	Disadvantages
SVM	• SVM is effective in high-dimensional space. • SVM is a relatively memory systematic • SVM can model nonlinearly separable data by using the kernel trick	• SVM does not perform well when the dataset has noise. • SVM will underperform if the number of features exceeds the number of data samples.
MLP	• Can easily work with nonlinearly separable data. • MLP is relatively easy to train • It is more robust to noise.	• Too many parameters need to be optimized. • The high number of fully connected nodes results redundancy and inefficiency. • Large NN is Computationally costly.
PNN	• Much faster. • More accurate compared to MLP and SVM. • Very robust to noise.	• Requires much more memory space. • Slower than MLP in case of classifying new classes

Chapter 3. Combined machine learning techniques for analyzing the back contact influence on the stability of perovskite solar cells

3.1 Introduction

Even though the power conversion efficiency of the perovskite solar cells has achieved significant progress in recent years. Nevertheless, the problem of the stability degradation of the device remains a dilemma, this degradation occurs mainly due to the contact of the device layers with H_2O [60]. Therefore, many researchers have tried to mitigate this phenomenon through different approaches. In this context, improving the stability of the electrode contacts can enhance the overall device stability. This enhancement can be done by the adoption of different material compositions for the back and front contacts [61]. However, most of the research that adopts this approach are using the trial-and-error method [29]. This means that this research relies on the production of a variety of PSC devices with different electrode contact components, and then measuring the extent of the stability degradation of these devices. However, this approach is very expensive and time-consuming due to the large options of materials [29]. Furthermore, many research papers provide different ML approaches for predicting stability using new materials [62, 63, 64].

Recently, many researchers have shifted to using ML techniques in the design of solar cells and resolving the stability problem. For instance, J. Schmidt et all used a number of machine learning techniques with Density Functional Theory (DFT) to predict the thermodynamic stability of perovskite materials [65]. Ç. Odabaşı and R. Yıldırım analyzed different back contact components of 404 PSCs samples by using the Decision trees and association rules Apriori algorithms [60]. However, using a single decision tree cannot fit heterogeneous data and mostly leads to overfitting problem [66].

Therefore, in this chapter, we have used two ML techniques of eXtreme Gradient boosting (XGBoost) and MLP, we have applied the XGBoost algorithm for analyzing the effect of the back contact on the stability, along with MLP Regressor to predict the PSC device stability with different back contact components. This chapter aims to determine which is the best material employed in the back contact to enhance the PSC device stability.

In particular, this chapter is organized as follows: we will first explore the XGBoost technique and describe the dataset used in this chapter. Then, the preprocessing stage and hyperparameters tuning are revealed. Further, the obtained results are presented in the Results and Discussion section. Finally, we have provided a recommendation for which back contact compounds can enhance the PSC device stability, and finish with a conclusion.

3.2 Materials and Methods

3.2.1 Extreme Gradient Boosting (XGBoost)

eXtreme Gradient Boosting is considered a framework that implements the Gradient Boosting algorithm [67]. However, XGBoost is a scalable end-to-end tree boosting system available as an open-source package in the following link: https://github.com/dmlc/xgboost. Which distributed a Gradient-Boosted Decision Tree (GBDT) machine learning library. Basically, XGBoost is based on supervised machine learning which can be used for regression, classification, and ranking problems.

Moreover, XGBoost is one of three ensemble learning algorithms used in this research along with Extra Trees and Random Forest. The ensemble learning algorithms combine multiple algorithm predictions to obtain a better model. XGBoost, ET, and RF all consist of multiple decision trees, the difference is in how the trees are built and the predictions are made. Figure 3.1 shows the difference between an ensemble tree algorithm and a single decision tree algorithm.

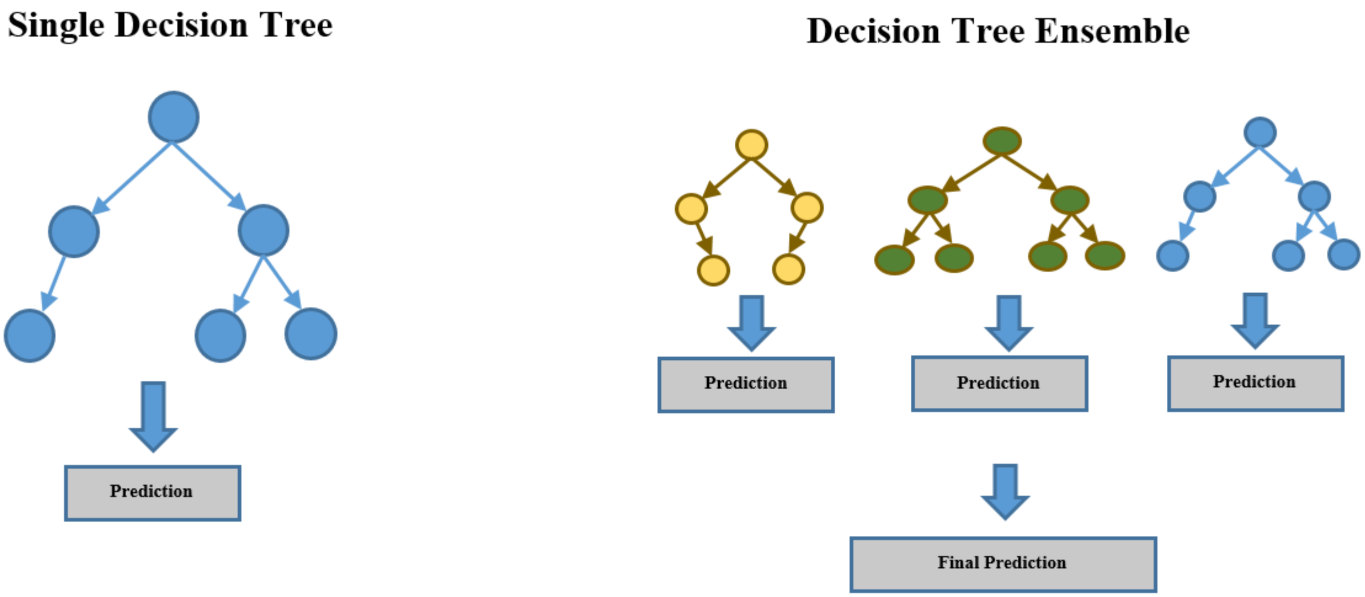


Figure 3.1- Difference between a single decision tree algorithm and a decision tree ensemble algorithm.

In the case of tree-boosting algorithms, the final prediction is calculated as the sum of the predictions for each tree. For example, given a dataset contains n examples with m features {(x_i, y_i)}, the model uses the sum of k functions to predict the output as described in the following equation:

$$prediction_i = \sum_{k=1}^{k} f_k(x_i) \quad (9)$$

Each f_k corresponds to a specific tree structure.

However, XGBoost has achieved significant results in many applications. For instance, Kaggle published the results of a machine learning competition in 2015. Interestingly, among the 29 winning solutions, 17 solutions used XGBoost. Moreover, among these solutions, nine of them used XGBoost combined with neural networks algorithm [67]. An example of problems solved in this event: high energy physics event classification, and product categorization.

3.2.2 Dataset Description

The dataset used in this chapter contains 140 different material configurations of a regular structure PSC. Also, it provides the stability T80 for each device collected from previous experimental research. This dataset was collected manually by C.Odabas and R. Yildirim [60]. However, in the next chapter, we will see that the environmental conditions of the PSC devices are very effective on the device stability degradation if exceed the ambient conditions. Therefore, we have already deleted the samples that are stored in extreme environmental conditions. This criteria could certainly neutralize the effect of the temperature and humidity on the device stability which allows to analyze only the influence of the materials involved in the back contact.

The ML model inputs are the materials constituted by the different PSC layers. Where the output is the stability for which the cell preserved 80% of its initial PCE.

3.2.3 Data preprocessing

In the material dataset, the manual exploration of the data is a long process but it is necessary to organize the dataset and to correct the typing mistakes. For example, if the same component of the perovskite active layer was written differently (e.g.: $FAMAPbI_2$ with FA-$MAPbI_2$), in this case, the ML algorithm will classify it as two different components. Therefore, each category must be written in the same way. Also, an extra space which often overlooked causes categorical mistakes. Hence, the material datasets should be well examined for ML application.

However, to ensure the effectiveness of the data, a different preprocessing method was used. The categorical features were converted to numerical values by using the scikit learn library of LabelEncoder. This method encodes a categorical label into a value between 0 and n – 1, where n is the number of the categorical labels. However, Scikit Learn announced that this method should be used to encode only the target values, i.e. encode y rather than the input x. Nevertheless, using this method to encode the inputs in this chapter solves the problem of the categorical features without affecting the performance of the model heavily. It is also widely used for this purpose. We avoided using OneHotEncoder due to the large number of categories existing in this dataset, which will significantly increase the dimension of the dataset.

Finally, we have divided the dataset into two subsets. A subset contains 80% of the total data to train the machine learning models. The remaining data was used to test the models.

3.2.4 Solution approach

The objective of this chapter is to enhance the PSC device stability by finding the appropriate material component of the back contact layer. Therefore, we have adopted two approaches:

First, we have investigated the influence of the back contact on the total stability of the PSC device. Hence, we can evaluate the improvement of the device stability that can be obtained by using adequate material in the back contact. This process was achieved by using the feature importance technique from XGBoost. This technique is one of the benefits of using ensembles of decision trees, which can estimate how the importance of a feature is relative to the target (in this case, the device stability). However, the importance of the feature importance technique can be calculated with three different importance metric as follow [68]:

-The coverage metric: is the number of observations related to a specific feature divided by the total observations. For example, we have a set of observations and 3 decision trees, the input contains 10 features, and suppose one feature is used to form a leaf node with 10, 8, and 5 observations for tree1, tree2, and tree3 respectively. Then the cover metric will calculated for this feature as follows: 10+8+5=23, in the same way the cover metric of the other features will be calculated. The importance score is the percentage of 23 overall feature cover metrics.

-The frequency/weight: is the percentage of the number of the appearance of a specific feature in all model trees. For example, if a feature occurs in 5 splits in tree 1, and 2 splits in tree 2, the importance is calculated as the total weight of this feature overweights all features: $\text{importance} = \frac{5+2}{total\ weight}$

-The gain: is the relative contribution of a specific feature to the model, it's calculated by taking the contribution of each feature for the model trees, as shown in the following equation:

Breiman et all proposed [69], for a single decision tree T:

$$T_l^2(T) = \sum_{t=1}^{j-1} i_t^2\, I(v(t) = l) \qquad (10)$$

The sum is calculated over j-1 nodes of the tree. For a t node, one of the input features is associated with splitting that node into two subregions (nodes or leaf), the chosen feature is the one that gives maximal estimated improvement. The XGBoost model generalizes the method above across all the trees used in the ensemble and then calculates the average. However, this type of feature importance is the most indicative of the contribution of a feature relative to the target. Hence, we have used the gain type of feature importance.

The second approach is based on choosing a specific PSC structure, and then predicting the stability of this device several times while changing the back contact component. The back contact material of the highest device stability is considered an adequate candidate for enhancing the device stability. However, to achieve more reliable results, we have repeated this process with five different PSC structures. This operation was done by using the MLPRegressor algorithm.

3.3 Results and Discussion

The evaluation of the performance of the two models was calculated over the test set by using the cross-validation technique from scikit learn, the method for estimating the accuracy is MSE. Furthermore, the XGBoost accuracy is 70.6%. The evaluation of the MLPRegressor algorithm gives 85.39% of accuracy. However, this increment of accuracy between XGBoost and MLPRegressor models is due to the type of the dataset used, the material dataset is considered highly heterogeneous which gives the advantage to the neural networks algorithm. Moreover, we have traced the learning curve of each algorithm in Figure 3.2 by using the Learning Curve class from the Scikit Learn library. Which is a diagnostic tool in ML that represents the model performance changes during the learning and test process by using cross-validation to split the training set 5 times, and the score of each subset was computed.

Figure 3.2-a shows that the training score is relatively much higher than the test score, which means that the XGBoost model overfits the training data. In particular, it is expected due to the fact that the pattern in our dataset is too complex which increases the variance. However, the curve test shows a progressive increment in terms of accuracy whenever the number of data samples used to train the model increases. This means that the larger the dataset is, the better the model performance. Moreover, the overfitting phenomenon is a problem for a predictive study of new materials, but it is unlikely to affect the feature importance technique because it is based on computing the training data.

From Figure 3.2-b, we can conclude that the MLPRegressor is pursuing a good learning process, in which the accuracy of the training set is relatively close to the accuracy of the test set. In general, after the training process, the results of feature importance are shown in Figure 3.3, which indicates the influence of each PSC layer on the total device stability.

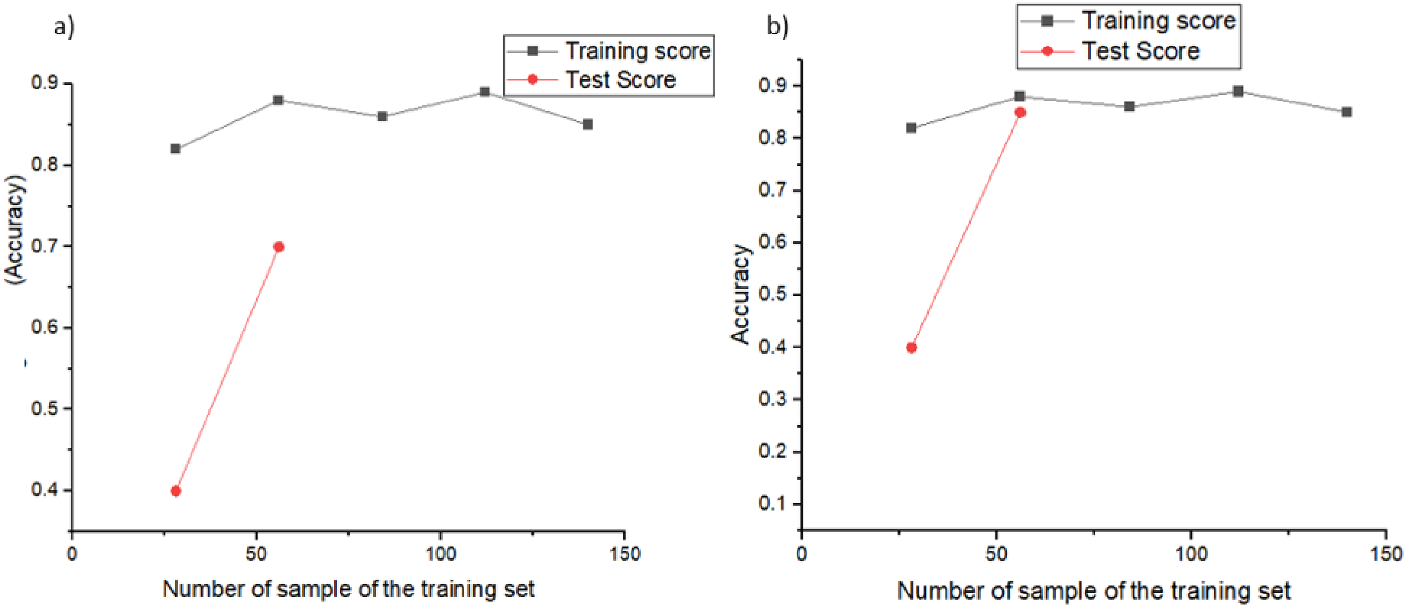


Figure 3.2- Learning curve of: a) XGBoost. b) MLPRegressor

Figure 3.3 shows that the electron transport layer (ETL) affects the device stability by 19%, with the added ETL second layer, the effect becomes 33.8% (19% from ETL plus 14.8% of ETL-

2) which is a significant value. Nevertheless, we will investigate the influence of all different layers in the next chapter by using a more decent dataset with a different approach. However, the influence of the perovskite active layer the hole transport layer (HTL), and the HTL additive layer are 17.5%, 19%, and 14.7% respectively. The back contact affects 15% of the total device stability. In accordance, many experimental studies show that using different back contact materials leads to a significant change in the device stability. For example, F. Behrouznejad et al. proved that by using different materials as Back contact, Platinum (Pt) provides better stability compared to other materials like Silver (Ag), Copper (Cu), Nickel (Ni), and Chromium (Cr), in fact, the device shows improvement in the performance and stability due to the relatively high work function of Pt [70]. Moreover, Farhadi et all found that using the metal component as back contact minimizes the layer defects and the influence of the temperature on the PSC performance, which helps to improve the electrical characteristics and the stability of the device [71]. Similarly, Ç. Odabaşı and R. Yıldırım have found by using the ML techniques of association rule mining that the back contact is an important factor concerning stability [60].

Figure 3.4, shows the prediction of the stability for five different PSC devices, each device has a different perovskite structure, for example; multianion perovskite ($MAPbI_{3-x}Cl_x$), 2D/3D perovskite ($CsPbI_3$-$EDAPbI_4$), etc. Furthermore, the stability of each device was predicted by using six different back contact materials. The back contact materials used in this study are: Silver (Ag), Silver-Aluminum (Ag-Al), Gold (Au), Multi-Wallet Carbon Nanotube (MWCNT), Carbon, and Graphene. These are all the back contacts available in the dataset.

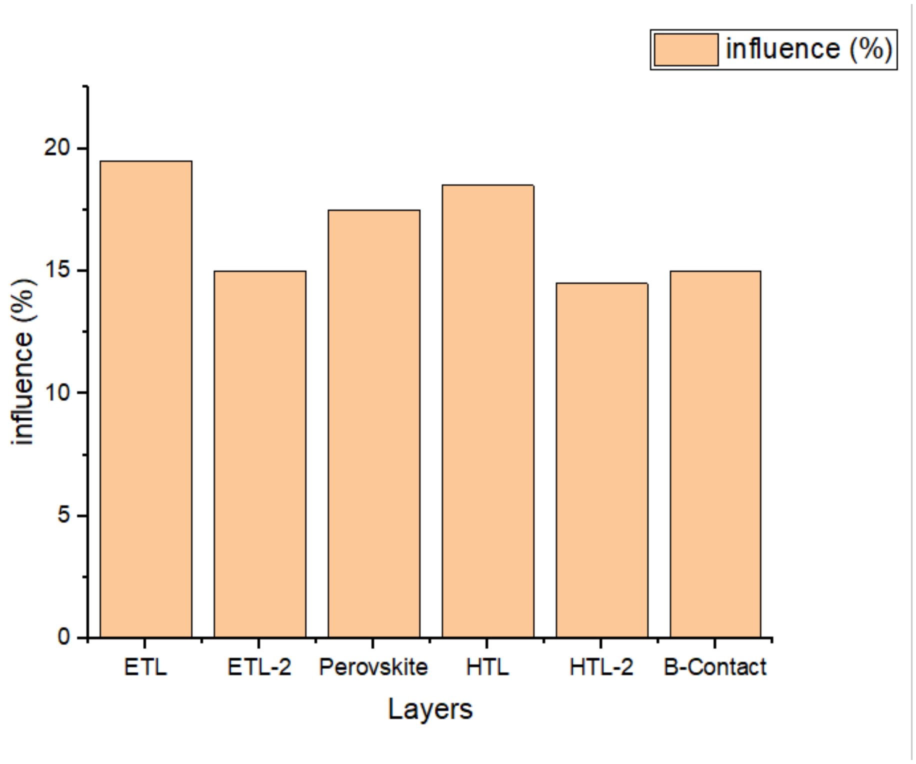


Figure 3.3- Percentage of the important score of the PSC layers

Figure 3.4 clearly shows that the stability increases when using MWCNT, carbon, and graphene. However, all these components are carbon allotropic forms. The reason for the enhancement of the device stability when using carbon alternatives may be due to the nature of the carbon, which is a hydrophobic material. In particular, the H_2O strongly decreases the device stability mainly due to the presence of a liquid electrolyte [46, 47, 60]. Hence, using carbon as a back contact may protect the different interlayers from the risk of exposure to the water [72].

Moreover, figure 3.4 shows that Au gives a high stability for the $MAPbI_3$ structure with spiro-MeOTAD as HTL, the predicted stability value is 358 days. Indeed, by searching nearly the same structure, the real experimental value found in the dataset was 360 days. However, this finding cannot be generalized because it is only a single case (shown in the green curve), while the other cases did not show a good stability. In particular, F. Behrouznejad et al. demonstrated that Au is the most suitable metal for use with spiro-OMeTAD [69].

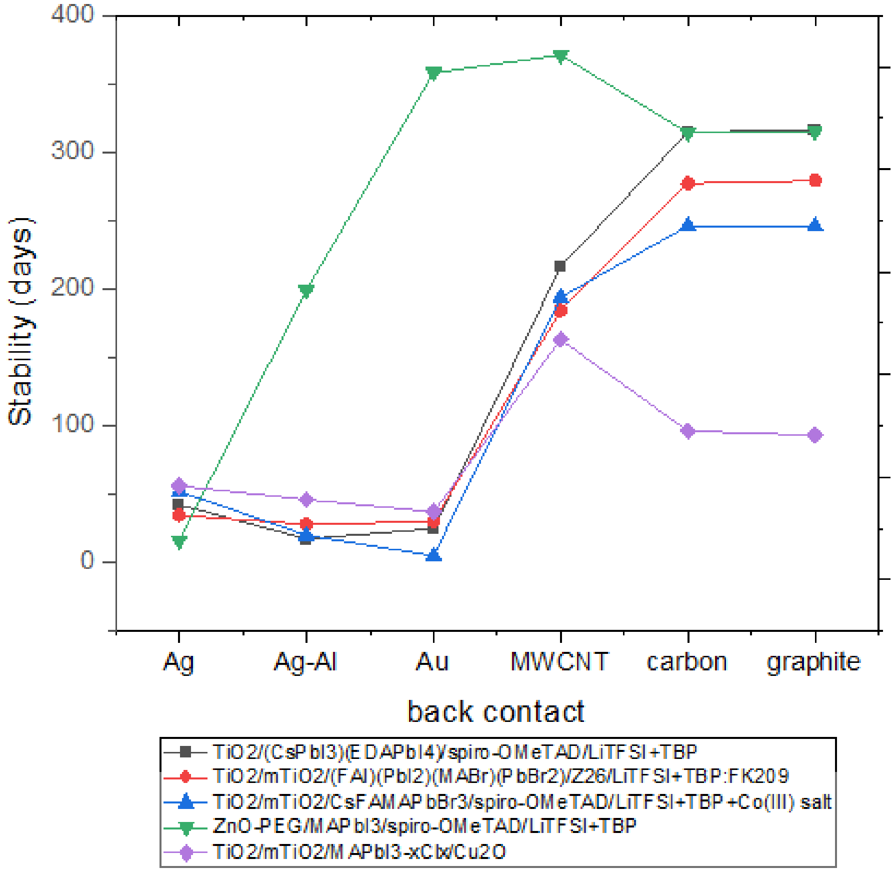


Figure 3.4 –MLPRegressor prediction for different material compounds of back contact.

3.4 Conclusion

In this chapter, we applied two different ML techniques of XGBoost and MLP on 140 experimental data samples of PSC devices. Our objective was concerned with enhancing the PSC device stability by using an adequate material as a back contact electrode. The XGBoost algorithm was used to estimate the influence of different PSC layers on the total device stability. Where the

MLPRegressor algorithm was used to predict the stability of different PSC devices with different back contact components. The results of both algorithms have been validated by using previous experimental published works. We have found that the back contact materials that are robust to environmental conditions significantly enhance the device's stability. Further in this research, the proposed approach will be enriched with more experimental data. Furthermore, it is possible to examine more layers with the aim of optimizing the entire device structure.

Chapter 4. Machine learning solutions for perovskite solar cells stability enhancement

4.1 Introduction

The stability degradation of the PSC remains the main issue that prevents the successful commercialization of this solar cell technology. Therefore, in this chapter, we intend to investigate the different factors that affect the device stability by using ML techniques of Extra Trees (ET) in an attempt to understand the degradation process of PSCs under different conditions and propose a solution to achieve high long-term stability for the PSC devices. Furthermore, the ET algorithm is used to analyze a large set of experimental data with 1050 data samples containing the material compositions of the PSC, the deposition methods, deposition solutions, and the environmental conditions. However, we avoided using neural network algorithms in this study due to their nature, which makes a quite impossible to understand the reason behind the outcome of these algorithms, and are not interpretable. Conversely, the algorithms based on decision tree gives many information that helps to understand the data. Moreover, the ET is usually preferable -in this case- compared to the previous techniques, which use the majority voting of decision trees to produce the final outcome, resulting much faster model, also the randomization of this algorithm decreases the variance of the model which eliminate the overfitting [65, 73]. However, due to the vast different factors involved in the PSCs manufacturing process, we are employed to analyze the most important factors related to the stability. The features that are relevant to the device stability were extracted using the feature importance algorithm from ET, while the different factors were investigated by using the ET classifier algorithm.

This work is structured as follows: at first, we will explore the dataset. Then, we are going to identify the Extra Trees technique and show the various steps of the preprocessing stage. Further, we will investigate the effect of environmental conditions on the device's stability by taking samples from the dataset. The results of the machine learning will be analyzed and compared with previous experimental works in the results and discussion section. Finally, we will propose an optimized device structure and predict its stability using an ET regressor, and finish with a conclusion.

4.2 Materials and methods

4.2.1 Dataset Construction

The dataset used in this chapter was collected from previous works by reviewing experimental data about PSC stability. However, the data was collected manually respecting several criteria: the cells under extreme environmental conditions like strong light intensity or high temperature were ignored, also the data should provide the stability T80 of the PSC device, the data should contain the different layer components, the different technique used in the device production, and the environmental storage conditions. Most of the dataset was collected from reviews articles [29, 60, 74,75,76], and the Perovskite Database Project [51]. Furthermore, this dataset contains 30 different features related to the device manufacturing and proprieties including: the cell architecture, the ETL and ETL second layer components and deposition procedure, the ETL thickness, the perovskite composition and thickness, the perovskite deposition steps and procedures, the deposition solvent and quenching media, the HTL and HTL-2 component and thicknesses, the HTL deposition solvent and procedure, the back contact component and thickness, storage light intensity, humidity, and atmosphere. Compared to the previous chapter's dataset, this dataset contains much extended number of features which helps to cover the majority of the factors that affect the device stability. Moreover, these features are the input for the ML algorithm, while the output is the stability of the cell.

4.2.2 Extra Trees

Extremely Randomized Tree (Extra Tree) is a tree-based ensemble method for supervised machine learning that can be used for classification and regression problems. The implementation of this technique was given by Pierre Geurts et all in 2006 in their paper entitled "*Extremely randomized tree".* Despite that, the main implementation for this technique in this paper was to process numerical values, although it can be adapted to process categorical values. The trees created by this technique are totally randomized and their structure is independent from the target values of the training samples. i.e. an ensemble of unpruned (de-correlated) decision trees [73].

However, the two main differences between ET and the other techniques that are the ET uses all the learning samples to grow the trees, and it uses the cut-points totally at random to split the nodes [73] which has a significant variance reduction effect (degrease the model overfitting). The final prediction is calculated by the majority vote in case of a classification problem, and the arithmetic average in case of a regression problem.

In particular, from the paper that introduced the ET model, all the 12 problems analyzed using different tree based models show that the ET has a lower variance (less overfitting) but a relatively high bias. The paper explains that due to the randomization in the algorithm which includes the irrelevant features in the model. Therefore, we intend to exclude the irrelevant features during the preprocessing step to ensure that we obtain the best performance of the model. However, the ET is considered an "in the clear" algorithm, which means that all the computations are done in the form of plaintext and clear operations rather than a black box which is the case in the fully connected nodes in the neural networks algorithms. So, let's have a little bit of a sense of how this method works, as this will help us to understand the relationship between our data and the stability of the device.

Considering the number of numerical input variables and two categorical target variables, the ET algorithm will build the DTs as follow:

Step 1: pick a random split

To build a decision tree, the algorithm chooses randomly several features without replacement at each node, the number of features picked is denoted k, and the minimum sample size for splitting a node is denoted n_{min}, for example: a dataset (S) contains numerical variables with A_i attribute, i∈ $\{1, 2, \dots., N\}$, and two classes red and blue.

A_1	A_2	…	A_i	Target
1.8	1.5	…	1.0	red
2.1	2.8	…	3.6	blue
4.0	1.1	…	1.3	blue
2.2	3.9	…	2.8	red
3.1	1.7	…	4.1	red
1.2	2.9	…	3.1	blue

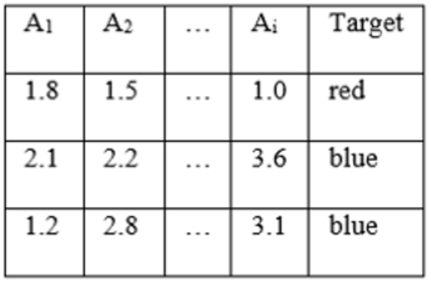
Suppose that k=2, and the algorithm chooses A_1 and A_2. Then, compute the max and min values in each feature in S, denoted a_{min}, a_{max} respectively. Draw a cut-point (threshold) a_c between [a_{min}, a_{max}], in our example, let's suppose that the thresholds are 2.2 in A_1, and 2.8 in A_2. Then return a split [$a<a_c$].

Step 2: Build an Extra Tree:

The stopping condition: the algorithm will return an output –build a leaf- (class frequencies in case of classification or average output in case of regression) if the subset is smaller than the minimum sample size n_{min}, or if the learning samples or the output variables are constant in S. Otherwise, the algorithm will split S into two subsets (denoted S_l and S_r respectively) based on the cut-point. Our example S becomes:

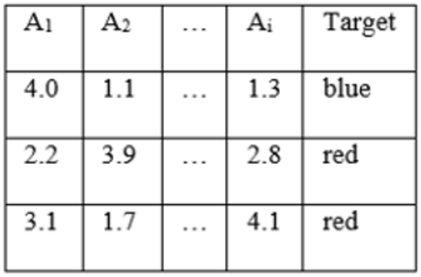
For the A_1 cut-point:

S_l (<2.2)



A_1	A_2	…	A_i	Target
1.8	1.5	…	1.0	red
2.1	2.2	…	3.6	blue
1.2	2.8	…	3.1	blue

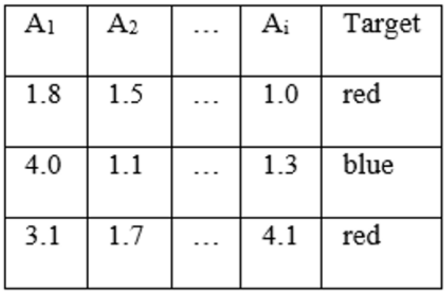
Sr (>=2.2)



A_1	A_2	…	A_i	Target
4.0	1.1	…	1.3	blue
2.2	3.9	…	2.8	red
3.1	1.7	…	4.1	red

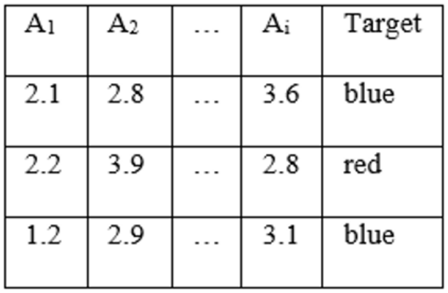
For A_2 cut-point:

S_l(<2.8)



A_1	A_2	…	A_i	Target
1.8	1.5	…	1.0	red
4.0	1.1	…	1.3	blue
3.1	1.7	…	4.1	red

S_r(>=2.8)



A_1	A_2	…	A_i	Target
2.1	2.8	…	3.6	blue
2.2	3.9	…	2.8	red
1.2	2.9	…	3.1	blue

Note that in the first split, the dataset contains only the rows where A_1 values are less than the threshold (2.2), the same thing in A_2 split. However, the subsets aren't required to be the same size. Now the algorithm will compute the best feature split that describes the data by using a quantifiable metric, there are several methods to calculate the score of how meaningful the split is in the dataset. In this work, we have chosen the Gini index.

The Gini index represents the probability of a sample picked randomly to be misclassified. This means that the lower the Gini index is, the lower the chance of an instance being incorrectly classified. Moreover, the value of the Gini index is between 0 and 1. The formula of the Ginin index can be written as follows:

$$Gini = 1 - \sum_{j=1}^{j} P(j)^2 \qquad (11)$$

j represents the number of the class in the target variable (j=2 in our example). Where P(j) represents the ratio of the pass/total number of observations (learning samples) in the node. Given this definition, the weighted sum of the Ginin index can be calculated as follows:

$$Gini = \ 1 - \sum_{1}^{2}(\frac{S_*}{S} \ (\sum_{1}^{j} P_{*,J}^2)) \qquad (12)$$

S_* represent the two subsets {S_l, S_r}. Therefore, the sum is from 1 to 2. Let’s calculate the gini index for our example:

$$Gini(A_1) = 1 - (\frac{3}{6} \ ((\frac{1}{3})^2 + (\frac{2}{3})^2 + \frac{3}{6}((\frac{2}{3})^2 + \left(\frac{1}{3})^2\right) = 0.45$$

In this example, the subnode has 3Pass and 3Fail, in the 3Pass samples, the number of red class is 1/3, and the blue class is 2/3. Suppose that the same calculation was done for A_2 and gives the Gini index = 0.55. In this case, the ET algorithm will consider the split 2.2 of the feature A_1 as the best description for the data, and create the child node based on the subsets created from this split. The right child node has S_r subset, and the left child node has S_l subset.

The ET classifies the new data by comparing the current sample value in the A_1 feature with the cut-point, then this sample travers the sub-tree depending on its A_1 value in which the subset belongs, and then it continues recursively to fulfill all the similar criteria till reaching the stopping condition an build a leaf node. However, in this example, the ET algorithm will compute the majority vote of all the decision trees to produce the final classification. The pseudo-code of the ET algorithm is described in Table 3.1

Table 3.1: Pseudo-code of the ET algorihtm

```
Input: a training set S with A features and n samples.
Output: an ensemble of trees with M tree, $T=t_1, …,t_M$
Build_Extra_Tree(S):
   If $S<n_{min}$ or sample are constant in S or feature values are constant in S:
      Return: a leaf labeled (by class frequency in case of classification, by average output in regression)
   Else:
      1-Select k features randomly without replacement in S.
      2-Generate k splits $S_*$, where $S_*$ = Pick_random_split(S, $A_i$)
      3-Select a split such has the best description of data using a scoring function
      4-Split S into two subsets $S_l$ and $S_r$ based on the previous test
      5-Build sub tree $t_l$ = Build_Extra_Tree($S_l$), $t_r$ =  Build_Extra_Tree($S_r$)
      6-Create a node that attaches $t_l$ and $t_r$ as the left and the right sub-tree of this node
      Return: the resulting tree t
Pick_random_split(S, $A_i$):
      1-Find $a_{min}$ and $a_{max}$ in S
      2-Choose randomly a cut-point within the range of [$a_{min}$, $a_{max}$]
       Return: the split [$a<a_c$]
```

4.2.3 Data Preprocessing

The preprocessing step is extremely important in this chapter because the model learning process is directly impacted by it. Constantly, we start with the manual exploration of the dataset to class the categorical data appropriately under different categories and delete the data that has many missing values. However, this dataset contains 30 features which may cause a reduction of the model performance both in speed -due to the high dimensionally- and in accuracy -due to the redundant and irrelevant features that may create bias- [77]. Therefore, we have used the feature selection method to reduce the number of features and delete the irrelevant ones. The advantages of this method are schematized in Figure 4.1.

However, feature selection is considered one of the major problems in ML. This method aims to select the most important and non-redundant features to use in the model learning process without losing information [78]. *Furthermore, it is used to* make the model easier to interpret by researchers, make training faster, decrease the impact of the high dimensionality curse, and enhance the compatibility of data with the model [79].

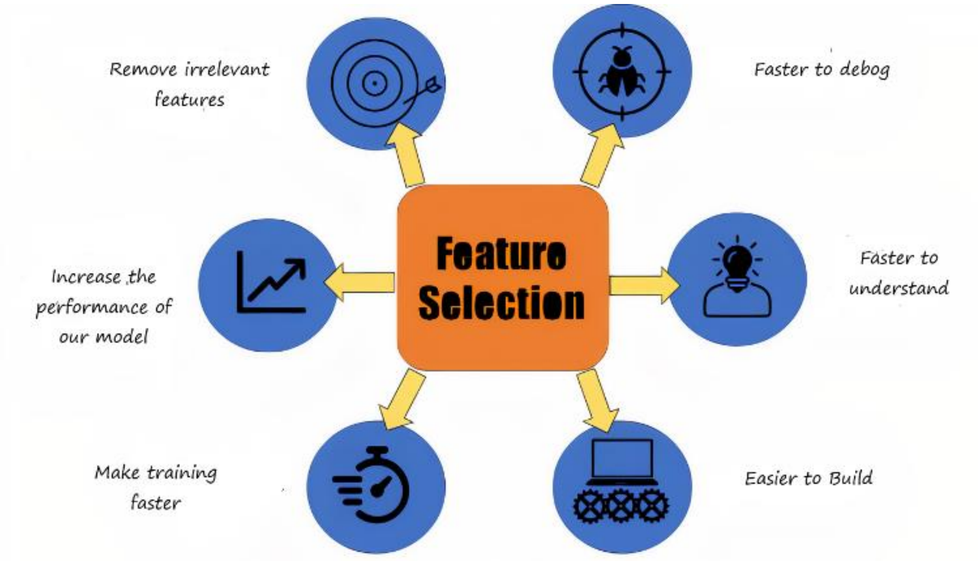


Figure 4.1-The advantages of the features selection process

In this chapter, this process relies on four steps: Encoding the data, Normalization, Imputation, and feature extraction, as shown in Figure 4.2.

The categorical data was encoded by using LabelEncoder from the Scikit Learn library. In particular, we used LabelEncoder just for feature selection, in the ML analyses we used OneHotEncoder to encode the categorical variable in order to transform the different PSC components into features to facilitate the analyses, we will discuss the approach used in this work further in this chapter. Moreover, to increase the learning speed and facilitate the convergence of the model we have used the normalization method to reduce the value of the encoded variable to a common scale. To retrain the maximum of the data, we used a simple imputation method from Scikit Learn to replace the missing data with an arbitrary category in case of a categorical feature or a constant number in case of a numerical feature.

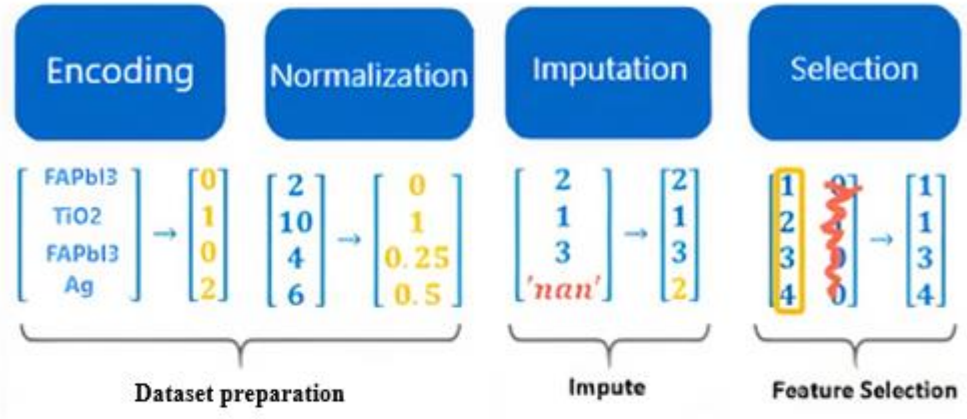


Figure 4.2- Feature selection process

The feature selection is considered supervised if the extracted features are selected based on the output variable, and unsupervised if not [80]. In particular, the feature selection can be applied by using two different methods; Wrapper methods, or Filter methods. The wrapper method consists of creating multiple models by using different subsets with removed features, this method keeps adding and removing features until finds the optimal combination, then choosing the features from the subset that result the best performance [80]. The filter method uses a statistical technique to compute a score that indicates the relevance between the features and the target and chooses only the features that fulfill some criterion. However, in this work, we have used the filter method by applying the Gain metric of importance score and choosing the features based on a threshold score.

Subsequently, we have divided the dataset into two datasets based on the structure of the PSC device, the first dataset contains 723 data samples of regular structure PSC, and the second contains 327 data samples of inverted structure PSC. Each one of these datasets was divided randomly into 80% of the data to train the model, and 20% of the data to test the performance of the model. However, we have used random train/test split to distribute the data uniformly and prevent model overfitting. As a result, the training data becomes a good predictor of the test data, and the test data becomes a good predictor of future data [81].

4.2.4 Machine learning approach for data analyses

The objective of this work is to investigate the effect of different factors on the PSC device stability, and then propose an optimized structure to achieve high long-term stability. Therefore, our approach is based on three phases: the first phase consists of extracting the relevant features related to the device stability and then create the ML model. Then, we investigated the effect of the different factors on the device stability by analyzing the decision trees built by the model and the importance score related to these factors. Finally, we proposed different configurations of PSC cells for both regular and inverted structures, then predicted their stability and compared it with the top experimental cells in terms of stability that are available in our dataset. Figure 4.3 illustrates the different steps of ML model building.

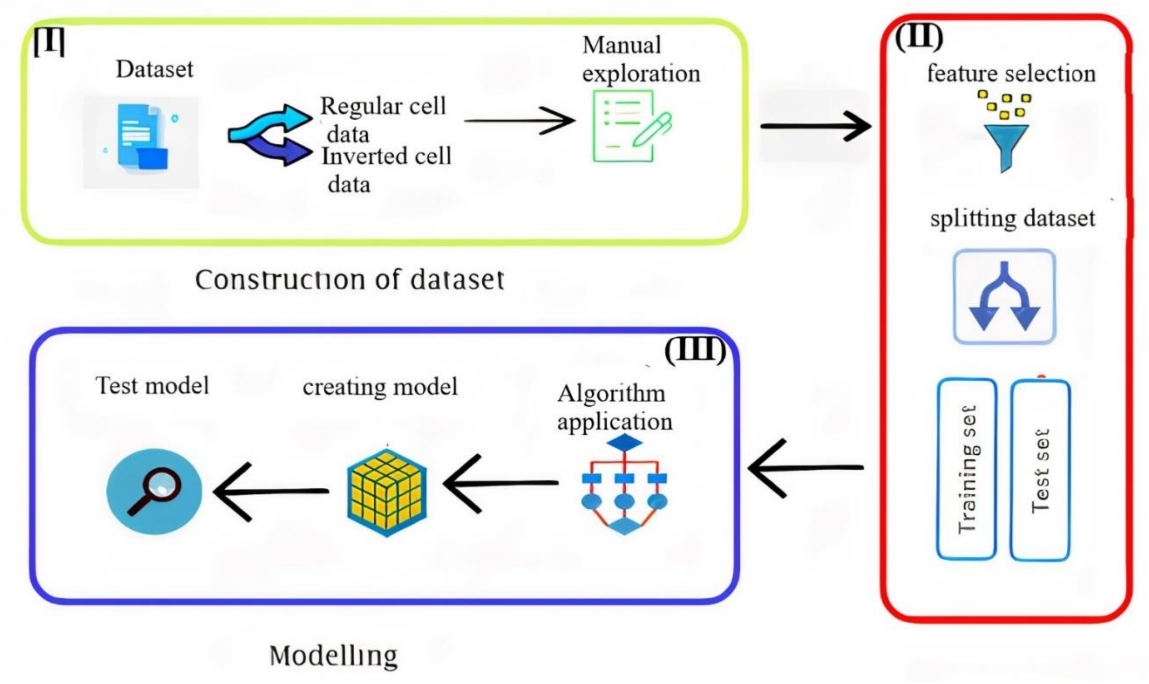


Figure 4.3- Different steps of creating the ET model

4.3 Results and Discussion

4.3.1 Analyze the Feature Selection results

The Extra Tree algorithm was used to apply the importance score method to exclude the factors that have a low effect on the device stability and select the most consistent features. However, all the features considered to be kept or removed from the dataset based on a threshold of the Gain metric method. The features were ranked based on their score, and the score where the next feature has a big drop in score value was considered as the threshold. However, the results of this method may change slightly due to the stochastic nature of this algorithm. After several processes, we were able to eliminate the irrelevant features. 20 features were removed and only 10 features were kept. The importance score of the remained features was also computed as shown in Figure 4.4.

However, to comprehend the feature selection results, we should first identify the different factors that impact the device's stability. In particular, the factors that determine the stability can be classified into two categories: the intrinsic and the extrinsic factors. Generally, the intrinsic stability degradation occurs during the stress of different operational conditions that lead to changes in the perovskite active materials proprieties. While the extrinsic stability degradation is related to environmental conditions such as humidity and temperature [82]. Based on the literature, the intrinsic stability highly depends on the ETL and HTL materials, the perovskite active layer (absorbent layer), and the back contact materials [70, 83]. Moreover, Aldibaja et al. found that the use of different lead precursors significantly affects the stability of the devices, and the utilization of the $PbCl_2$ as a precursor solution improves the device stability [84]. While Roghabadi et al. refer to the structural phase as an intrinsic factor as well as moisture, thermal, and light exposure as external factors that can urge the stability degradation of the PSC [85]. Moreover, the quick degradation of stability was found to be caused by heat and light [86, 87]. For instance, at 60^0C, the $MAPbX_3$ perovskite decomposes to gaseous methylamine, lead halide, and hydrogen halid [88].

Interestingly, the results of the feature selections show that the temperature, light condition, and oxygen have an inconsiderable effect on the device stability in this dataset, which seems a contradiction with the previous studies. However, these results may be justified due to the specific criteria during the data collection. The temperature feature in the dataset contains only the values close to ambient temperature. i.e 24 °C < T < 30 °C, which considered as almost constant. This means that it has a minimal estimated improvement in the splitting process of the node, hence, a very low gain score. Furthermore, from the previous chapter, the temperature causes a significant degradation of the stability when its value exceeds the ambient temperature. The oxygen-rich and light intensity show a low influence on the stability due to their low values in this dataset.

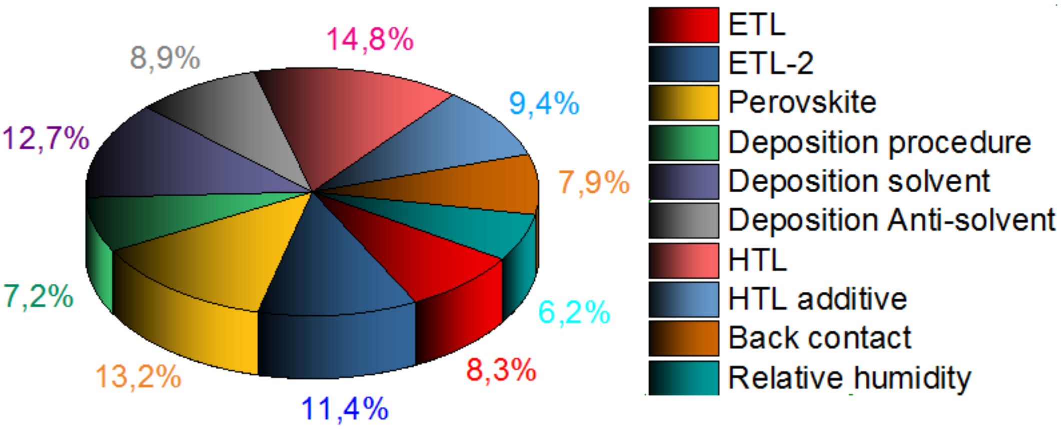


Figure 4.4- Influence of different factors on the perovskite solar cell stability

However, the relative humidity shows a significant importance score. Similar to the previous chapter, the relative humidity values were divided into three categories: 0-30%RH, 30-55%RH, and above 55% RH. This division is based on the results of the previous chapter and several previous studies. For instance, the cell stored under a relative humidity of less than 30% shows the same stability degradation for a fixed temperature at ambient conditions [59]. Furthermore, J.Noh et all found that the cells stored between 30% RH and 55% RH showed similar degradation pattern. And the cells stored under low humidity do not show any significant degradation [89]. Significant changes in the stability degradation were observed after exposure to 60% RH or above [59]. K.Ogunniran and N.Marins prove that the $MAPbI_3$ stability starts to degrade when exposed to 55% RH or above, otherwise it shows good stability [90]. However, Frost et all show that the decomposition of the $MAPbI_3$ caused by the humidity occurs in a reversible reaction as follows [91]:

$[(CH_3NH_3) + PbI_3]$ n + $H_2O \leftrightarrow [(CH_3NH3)$n-1+$(PbI_3)$ n]+$[H_3O] + CH_3NH_2$ React. 1

$[(CH_3NH_3)$n-1+(PbI_3)n]+$[H_3O] \leftrightarrow HI + PbI_2 + [(CH_3NH_3)PbI_3]$n-1 + H_2O React. 2

4.3.2 Extra Trees analyzes

The device's PCE and long-term stability are greatly influenced by the intrinsic factors. As we have seen in the previous chapter, using hydrophobic materials as back contact or ETL layer enhances the PSC stability. Moreover, using materials based on FA as an active layer provides more thermal stability due to the relatively large FA cation compared to the evaporative MA cation [92].

The importance score of the intrinsic factors within a specific feature was extracted. Each score of the materials under the same feature was compared. Hence, the most important materials for the device's stability were revealed. Besides, we have analyzed the decision trees built by this model to extract the stability classification of these factors. The objective of the combination of these two processes is to uncover the most important factors in each layer that may enhance the stability and give a recommendation for the best materials. An example of one DT is depicted in Figure 4.6.

The hyperparameters of the ET are as follow: the number of the trees is 310, the split criterion is the Ginin index, and the maximum depth of the tree is 3. Figure 4.6 shows an individual decision tree, the "sample" parameter indicate the number of samples that are used to split the node, as we have seen before, the ET algorithm uses all the training sample to split the first node. However, because the majority of the values of the device stability in our dataset are very low (sometimes less than one day), we turned this into a classification problem, the different classes used are indicated in the "class" parameter inside the leaf in figure 4.6. The "value" parameter denotes the distribution of the concerned samples for each class. For example, figure 4.6 shows that the samples with TiO_2 as ETL probably can lead to a high stability class.

Figure 4.7, 4.8, and 4.9 represents the different results information extracted from the ET algorithm, where the x-axes represent the different classes and the y-axes represents the importance score of the variable indicated by color. The bubble size represents the number of samples estimated in each class indicated in the x-axis. However, the figures don't show all the factors but only the promising variables. Moreover, figure 4.7 and 4.8 shows the influence of input variables for the regular structure PSC, and each graph represents a different feature.

Figure 4.7-a shows different materials used as an active layer, the materials that are found to be more stable are: $MAPbI_3$ with a 52% of importance score, 2D/3D structure combination with a 25% score, and Multi-cation perovskite with a 17% of score. Moreover, the inorganic CsPbBr shows relatively good stability. Several experimental results are consistent with these results.

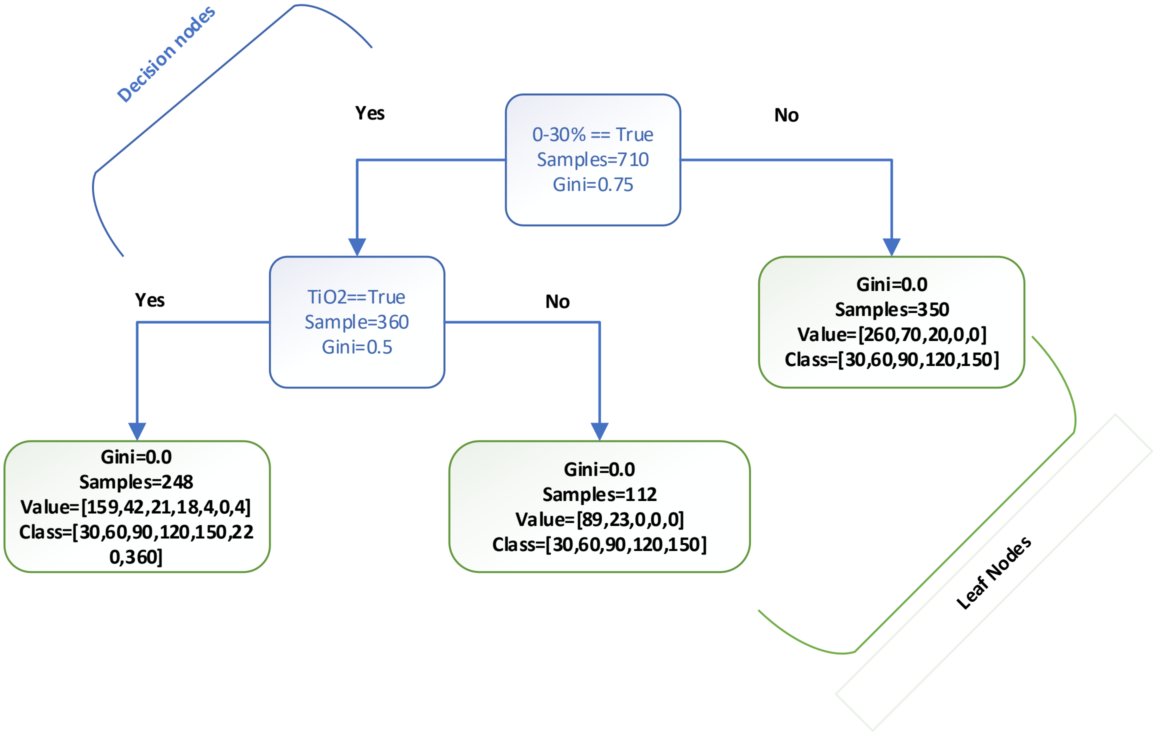


Figure 4.6- Illustration of a decision tree built by ET algorithm

For example, Gordello *et al.* prove that $MAPbI_3$ with a two-step deposition method enhances the stability of the device due to the improvement of their morphology associated with the increment of the grain size and with low density of structural defects [60]. Furthermore, a detailed study shows that the dissolution of the $MAPbI_3$ when exposed to water in darkness forms a molecular hydrate compound which improves the stability of the perovskite [90].

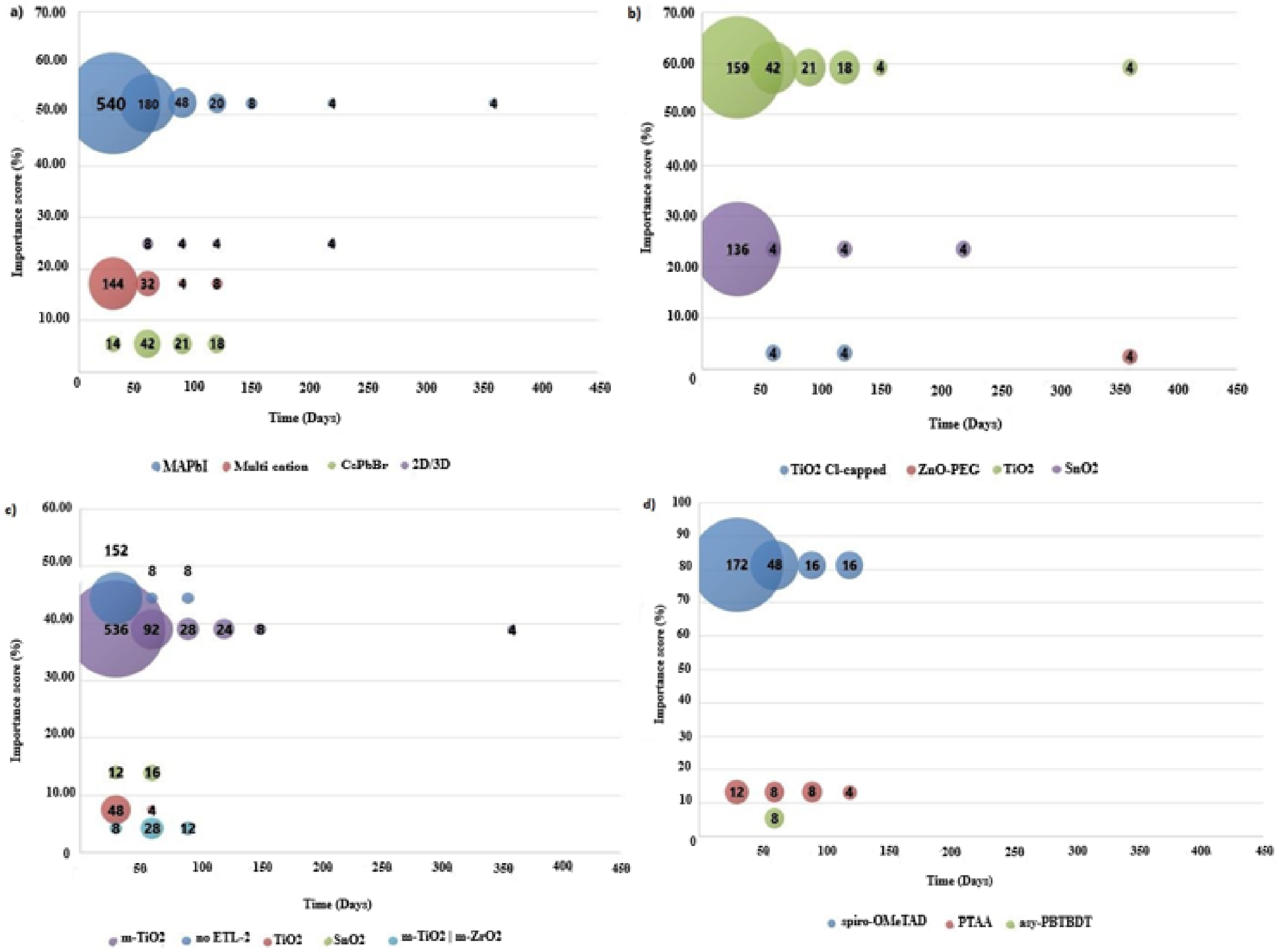


Figure 4.7-. Change of variable importance with the time classification probability for regular cells (n-i-p): (a)Perovskite active layer, (b) *ETL*, *(*c) *ETL-2*, *(*d) *HTL*

. The multi-cation perovskite is considered a relatively very stable perovskite. For instance, a multi-cation perovskite of Cs_x MA_{1-x} $Pb(5\text{-}AVA)_xI_{3-x}$, which was developed by incorporating the Cs(5-AVA) acetate with $MAPbI_3$ shows significant intrinsic stability, interestingly, it shows adequate stability when exposed to 100 ^{o}C of temperature for 500 hours and still maintain 88% of its initial PCE [93, 94]. The tolerance factor obtained from a multi-cation perovskite formed by mixing Cs and FA is between 0.9 and 1, which indicates that the perovskite crystal structure is very stable [59]. Moreover, different studies show that multi-cation perovskites with FAMA or CsFA have favorable stability at ambient conditions [95, 96].

Furthermore, from Figure 4.7-a, the 2D/3D structure of the perovskite has the highest ratio in high stability time, a 12 sample from 20 was classified with more than 90 days of stability. Fairly, the high stability of the multidimensional 2D/3D structure owing to the integrated 2D structure which is considered as a highly stable structure against environmental conditions [97]. Moreover, C.MA et all formed a 2D/3D perovskite with CA_2PbI_4/$MAPbI_xClx$ and shows a high device stability at a high humidity level of 63% ± 5% without encapsulation, compared to the individual 3D structure of $MAPbI_3xClx$, the obtained structure shows a significant improvement of stability against humidity [98]. Grancini et all developed a 2D/3D perovskite with $(HOOC(CH_2)_4NH_3)_2PbI_4/CH_3NH_3PbI_3$ and achieved a one-year of stability [99]. By adding a capping layer of 2D halide perovskite of PEA_2PbSnI on the top of 3D $(FASnI_3)_{0.6}$ $(MAPbI_3)_{0.4}$ perovskite thin film, Yuan et all obtained a high humidity stable PSC device [100]. However, due to the presence of the hydrophobic organic cation and the extremely dense packing structure in the 2D perovskite, the grain boundary is very reduced, causing less contact with moisture and oxygen, which improves the extrinsic stability in the 2D/3D perovskite [101].

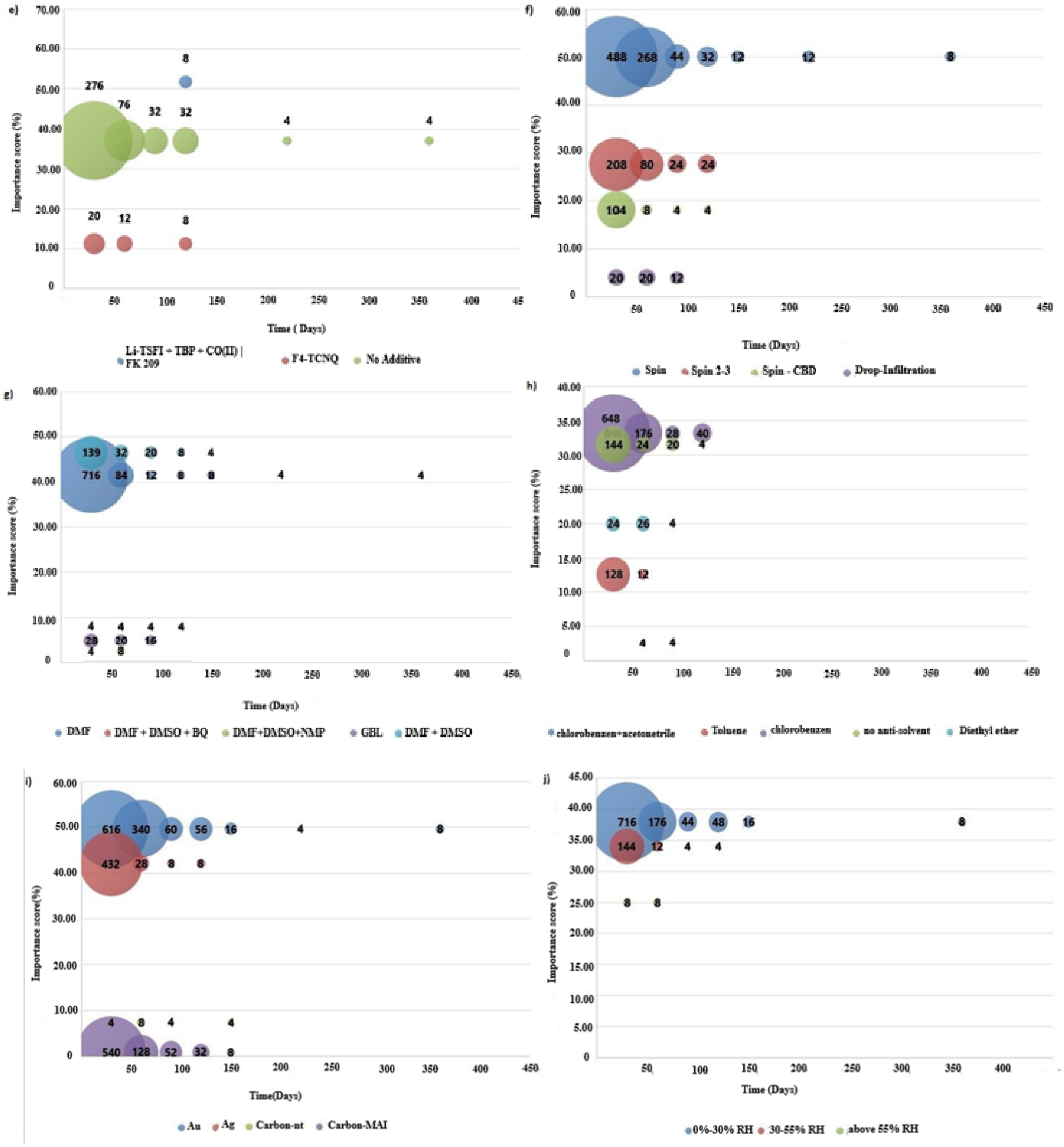


Figure 4.8- Change of variable importance with the time classification probability for regular cells (n-i-p): *(*e) *HTL additive (*f) *Deposition method, (g) precursor solution, (*h) *Anti-solvent solution, (*i) *Back contact, (*j) *Storage relative humidity.*

From Figure 4.7-b, the most effective material on regular cell stability is TiO_2 followed by SnO_2. Compared to all alternatives the TiO_2 has the highest appearance in the stable class. Furthermore, TiO_2 is the most used material as a photoanode in the PSC devices [102]. Although, it is considered as a typical photocatalyst, which leads to the oxidation of the organic cation of the perovskite [103]. However, during the exposure of TiO_2 to the light source, it starts to extract electrons from the perovskite materials that contain iodide as halide causing a deconstruction of the device structure under the following equation [104]:

$2I^- \leftrightarrow I_2 + 2e-$ [at the interface between TiO_2 and $CH_3NH_3PbI_3$] reaction.1

$3CH_3NH_3 + \leftrightarrow 3CH_3NH_2 \uparrow + 3H^+$ reaction. 2

$I^- + I_2 + 3H^+ + 2e- \leftrightarrow 3HI \uparrow$ reaction.3

The formed HI will evaporate immediately due to its low boiling point. In Figure 4.7-b, the TiO_2 doped with Cl shows a slight stability improvement. However, different studies show that the TiO_2 doped with Cl, Al, or Nb improves the intrinsic stability of the PSC device in static environmental conditions [105, 106]. M. Shahbazi and H.Wang state that a PSC device with TiO_2-ZrO_2 as a photoanode gave 1000 hours of high stability under light source of AM 1.5 and ambient temperature [72]. The devices with SnO_2 as a photoanode were found to be more stable than the devices with TiO_2 [107]. Moreover, K. Junu et all compared two PSC devices formed with $MAPbI_3$ with SnO_2 and TiO_2 as ETL for each, the results show that the device based on SnO_2 exceeded the TiO_2 device instability and in the election generation [108].

Figure 4.7-c shows that the TiO_2 mesoporous has a high effect on stability when used as the ETL second layer. In general, the mesoporous structure is considered more stable than the planar structure due to the small contact area with moisture [72]. A.Mei et all tested the stability of a multi-cation PSC with ZrO_2 as ETL and mesoporous TiO_2 under a simulated sunlight source

(AM 1.5), and the device showed excellent stability for more than 1000 hours [109]. From Figure 4.7-c, we can see that not using an additive layer of ETL also has a high impact on the device stability, but only 8 of 168 subset sample was classified in good stable class. This means that using m-TiO_2 is the best choice to enhance the device's stability.

Figure 4.7-d shows that the spiro-MeOTAD, PTAA, and asy-PBTBDT are the most promising materials for use as HTL. However, spiro-MeOTAD and PTAA are the most common materials used for this purpose [59]. However, the arylamine spiro-MeOTAD is considered highly-priced and may restrict the device's stability [109]. In our model classification, it has the highest impact on the stability and the lowest ratio of samples in the stable class, which means that it highly degrades the device's stability. Despite that, it has a great effect on the device efficiency, FK. Aldibaja et all formed a PSC device using a scaffold $MAPbI_3$ with TiO_2/m-TiO_2 as ETL and ETL-2, and spiro-MeOTAD with 300-400 nm of thickness as HTL, and Au as back contact, the results show that using spiro-MeOTAD as HTL leads to the degradation of the perovskite active layer stability [84]. By using a different configuration of PSC devices, M.Spalla et all found that the cells containing PTAA as HTL give a high performance but its stability significantly degraded under humidity due to defection of the PTAA material when reacting with HI during the annealing process [110].

Figure 4.8-e shows that Li-TSFI + TBP +Co(II) is the most effective and stable alternative for HTL additive, the Li-TSFI and PEDOT:PSS may cause the degradation of the PSV layer due to their ability to absorb the moisture [59, 109]. However, doping Li-TFSI + TBP on spiro-MeOTAD or P3HT can enhance the device conductivity and the stability of the device [111].

The deposition method affects the PSC stability due to its impact on the crystallinity of the perovskite [59]. Figure 4.8-f shows that the spin coating and spin 2/3 are the appropriate deposition methods for stability enhancement. Figure 4.8-g shows that the precursor solution has a high impact on the device stability. Furthermore, the morphology and the size of the perovskite crystal are very dependent on the precursor solution [72]. The results of the ET model show that DMF and DMF + DMSO are the optimizing solutions, with some DMF samples classified as very high stability. However, the residues of the DMF after its evaporation process can degrade the perovskite material [102]. However, it was discovered that using DMF + DMSO as a deposition precursor improves the device's stability [59]. From Figure 4.8-h, using chlorobenzene or diethyl ether as deposition quenching media can affect the stability positively.

Similar to the results of the previous chapter, figure 4.8-i shows that the optimum back contact materials for stability are: Carbon, Ag, and Au. And from Figure 4.8-j, the recommended RH is between 0% and 30%.

From Figure 4.9-a, the most adequate materials for high stability are $MAPbI_3$ and $MAPbI_{3-x}Cl_x$, although, $MAPbI_{3-x}Cl_x$ has the highest ration in the stable classes. Moreover, V. Trifiletti et all formed by *$MAPbI_{3-x}Cl_x$* an inverted structure PSC with PCBM as ETL and NiO as HTL, the device shows a good performance both in efficiency and stability [112]. Interestingly, the ML results show that both NiO and PCBM as ETL and HTL respectively appear in the highest stable classes (figure 4.9-b and c). In fact, because PCBM has a low conductivity, it is recommended to be deposited in a thin film [108]. For an inverted structure of PEDOT: PSS/$MAPbI_3$/PCBM configuration, Jeon et al found that the optimum thickness of PCBM is 55 nm [96]. However, from Figure 4.9-b, C60 has the highest ratio in the stable class. Interestingly, doping PCBM with C60 protects the device from extrinsic factors by reinforcing the surface morphology of the ETL layer, which gives much better stability than using PCBM only [113]. From Figure 4.9-c, PEDOT:PSS has a high impact on stability with a good appearance in the stable class, whereas NiO has the

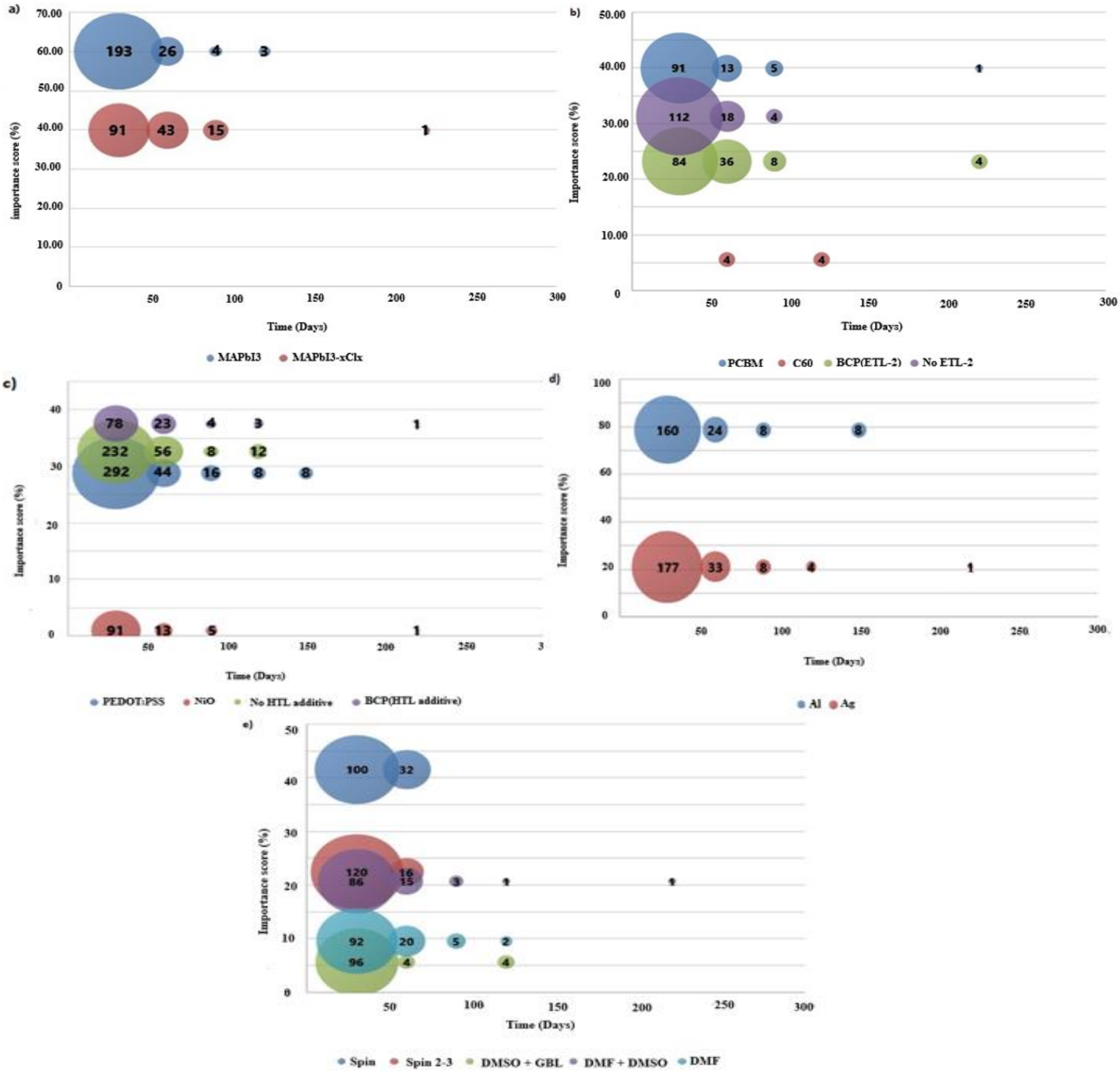


Figure 4.9- Change of variable importance with the time distribution for inverted cells (p-i-n): (a*)* Perovskite active layer, (b) *ETL and ETL-2*, *(*c) HTL and HTL additive, (d) Back contact,(e) Deposition method and precursor.

highest ratio in the stable classes. However, PEDOT:PSS is considered a hygroscopic which highly absorbs the water, resulting in a significant degradation of the device stability [72, 112].

Figure 4.9-d shows that the best back contact for the inverted structure is Al, where the optimum deposition method and precursor are spin and DMF + DMSO respectively, these results consistent with many researches [59, 112].

4.3.3 Materials and methods obtained by ET to enhance the device stability

In this section, we will propose an optimized structure configuration for both regular and inverted PSC based on the ET analyses, both configurations are illustrated in Figure 4.10- a and b. then, we will predict their stability and compare it with the top cell in our dataset in terms of stability in Table 3.2. However, the tree-based ensemble algorithm does not give predictions out of the target ranges used in the training process. Although, the predicted structures give a stability higher than 98% in the regular PSC case, and a stability higher than 93% in the inverted cells dataset.

However, the mesoporous structure was found to be more stable than the planar structure. Therefore, the proposed ETL contains TiO_2 and m-TiO_2 as the ETL second layer. The stable perovskite materials were found to have a 2D/3D multidimensional or multi-cation structure, $MAPbI_3$ is considered a good candidate for a regular n-i-p structure as well. However, the 2D/3D structure has the highest ratio in the stable classes. For the deposition method, spin and spin 2-3 are preferable for both regular and inverted structures, and DMF + DMSO and DMF as precursor solutions for regular and inverted cells respectively. Chlorobenzene was found to be a good deposition anti-solvent for both regular and inverted structures. As an HTL, despite that spiro-MeOTAD and PTAA have a high enhancement of PCE, their stability is poor. Therefore, the P3HT is considered a good candidate for highly stable cells. Moreover, the preferred back contact is the carbon due to its hydrophobic nature which protects the inter-layer from the humidity.

$MAPbI_{3x}Cl_x$ is the recommended perovskite material for the inverted structure. While PCBM doped with BCP or C60 was found to improve the device stability. NiO_x was found to be much better in terms of stability compared to PEDOT:PSS. The Al is the preferable back contact for the inverted cells. Finally, we illustrate the different results of previous similar studies in Table 3.3.

Table 3.2- Prediction of the optimized PSC structure compared to top experimental devices.

Cell configuration	*Deposition method*	*Anti-solvent traitement*	*Precursor solution*	*Structure*	*Stability time class*	*Reference*
TiO2/m-TiO2/(2D-3D)/P3HT/Carbon	*Spin*	*Chlorobenzene*	*DMF + DMSO*	*Regular*	*60+*	*This work*
TiO2/m-TiO2/(2D-3D)/Spiro-OMeTad/Au	*Spin 2-3*	*Chlorobenzene*	*DMF + DMSO*	*Regular*	*90+*	*[114]*
TiO2/m-TiO2/MAPbI3/Spiro-OMeTad/Au	*Spin 2-3*	*Chlorobenzene+acetonitrile*	*DMF+DMSO*	*Regular*	*60+*	*[115]*
NiO/MAPbI3xClx/PCBM/BCP/Al	*Spin 2-3*	*Chlorobenzene*	*DMF*	*Inverted*	*50+*	*This work*
NiO/DEA/MAPbI3xClx/PCBM/PN4N/Al	*Spin*	*No*	*DMF*	*Inverted*	*90+*	*[116]*
CuOx/MAPbI3/PCBM/Ag	*Spin*	*Chlorobenzene*	*DMF*	*Inverted*	*40+*	*[117]*

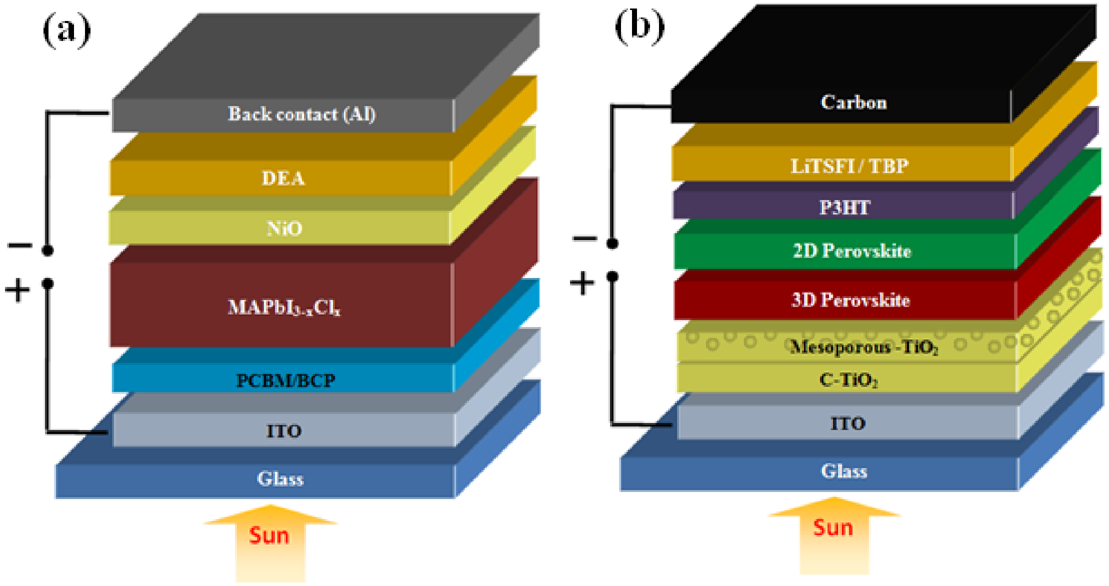


Figure 4.10- Recommended structure of: (a) perovskite regular cell, (b) perovskite inverted cell.

***Table 3.3**. Summarizing the results of intriguing similar works.*

Layer/Method	*Material*		*ML Technique*	*Ref*
	regular	*inverted*		
Perovskite	*(2D/3D)*	$MAPbI_{3-x}Cl_x$	*Extra Trees*	*this*
	Mixed Cation	*Mixed Cation (also* $MAPbI_{3-x}Cl_x$*)*	*Apriori algorithm, decision trees*	*[59]*
	NH_2NH_3InSI		*Gradient boosting regression (GBR)*	*[118]*
	$MAPbI_3$ *with PTEAI-capped* $((PTEA)_2(MA)_3Pb_4I_{13})$	*Mixed cation*	*Random forest regressor, association rule, Decision trees*	*[76]*
	Mixed cation	Mixed cation (also FA-based perovskites)	*Decision trees*	*[119]*
ETL / ETL 2	TiO_2 / m-TiO_2	PCBM / BCP	*Extratrees*	*This*
	SnO_2 / PCBM (also doped-mTiO_2)	PCBM + C_{60} / BCP	*Apriori algorithm, decision trees*	*[59]*
	SnO_2 (also doped-TiO_2)/ PCBM (also doped-mTiO_2)	PCBM + C_{60} / BCP	*Decision trees*	*[119]*
	TiO_2-dopped/m-TiO_2(or PCBM)		*Random forest regressor, association rule, Decision trees*	*[76]*
HTL / HTL additive	P3HT /LiTFSI+TBP+Co(II)|FK_{209}	NiO/DEA	*Extra trees*	*This*
	HTL-fre /F_4TCNQ	PTAA/ -	*Apriori algorithm, decision trees*	*[59]*
	Spiro-Ometad/ LTSFI+TBP(or LTSFI+TBP+CO(II)TSFI		*Random forest regressor, association rule, Decision trees*	*[76]*
	Inorganic HTLS / F_4TCNQ (also LiTFSI+ TBP + FK_{209})	PTAA (alsoNiO_x)		*[119]*
Back Contact	carbon	Al	*Extras trees*	*This*

	carbon	Cu (also Al)	*Apriori algorithm, decision trees*	*[56]*
	Carbon (also Ag)	Cu, Al	*Decision trees*	*[119]*

4.4 Conclusion

In this chapter, we have used machine learning algorithms of the Extra Tree classifier and importance score to investigate the impact of each PSC layer and deposition methods on the device's stability. We also analyzed the effect of the environmental conditions where the devices are operated or stored for both regular (n-i-p) and inverted (p-i-n) structures. Thereupon, the obtained results are compared with the state-of-the-art of previous experimental studies. However, we found that the environmental conditions are highly effective on the device stability principally the moisture. We found that the best materials to protect the device from moisture are those that have hydrophobic properties. Furthermore, the cells under high high-temperature environment were excluded. Interestingly, we found that the results of this work are compatible with many experimental results, which underpin the adopted approach. Correspondingly, in the next chapter, we attempt to optimize the PSC device structure to reach a better PCE.

Chapter 5. Perovskite Solar Cell optimization to achieve high Power Conversion Efficiency using Machine Learning techniques

5.1 Introduction

The PSC devices show a significant advantage in terms of the viability of efficiency increment, which has increased from 3.8% in 2009 to 26% in 2023 [16]. This rapid rise in PCE indicates that it has the potential for further improvement. Therefore, the objective of this chapter is to optimize the PSC device structure in an attempt to enhance the device PCE. The ML learning algorithm used in this study is Random Forest (RF) trained with a large dataset containing 3000 experimental data samples. The data was collected from several previous works involving the test of the efficiency of different PSC devices. Herein, these data contain many missing values that can reduce the ML model performance. Therefore, we have used different strategies to preserve the maximum of the data point. Furthermore, this chapter provides a guide for preparing a highly efficient PSC device by investigating the different factors that are related to PSC device manufacturing. Besides, the factors that have a significant impact on the device PCE were identified. The main factors investigated in this chapter are the materials and proprieties of the device layers (ETL, perovskite active layer, HTL, back contact) that were found to be efficient to the device PCE. However, the approach in this chapter is different from the previous stability analyses. To achieve high PSC PCE, we have to predict the efficiency of different PSC devices with different configurations and proprieties, and then, extract the values that are estimated to enhance the PCE.

This chapter is organized as follows: the next sections involve the different approaches used in this work. First, we provide the details of the ML technique used in this study, then, we present the data collection method and the strategy used to maintain the missing data. The preparation and preprocessing of the dataset for the ML application was also revealed. Further, we discussed the approaches used for the PSC PCE analysis and optimization. Then we will discuss the ML results and offer proposed structures and predict their efficiency. Finally, we conclude the results of this chapter and give future aspects.

5.2 Materials and Methods

5.2.1 Random Forest

Random forest (RF) is a frequently used ML algorithm developed by Leo Breiman and Adele Cutler in 2001 [120]. Similar to ET and XGBoost, Random Forest is considered as an ensemble learning method that uses a combining multiple decision tree predictors to improve the model performance and solve complex problems [120]. It is a supervised learning technique that can be used for both classification and regression tasks. Unlike the ET which builds de-correlated decision trees and uses all the data samples to build a tree, the generalization error of the RF depends on the correlation and the strength of the trees and uses a random data sample from the training data to build the trees. Moreover, the accuracy of the RF model increases (The cost function converges to a limit) with increasing the number of decision trees in the forest [120]. The RF algorithm comprised a forest of trees composed of random data samples and features bagging (i.e. features randomness) from the training data with replacement, this method is called the bootstrap sample. For a classification task, the final prediction consists of the majority vote of the decision trees in the forest. For the regression task, the final prediction is the average results of the individual decision trees in the forest. Herein, we are not diving deeper into this algorithm since we are only interested in the final output of the RF model. However, the reason for using RF instead of ET is due to the fact that we found that RF is much faster than ET (may be due to that RF uses fewer data samples to build the tree) since there are thousands of materials and proprieties prediction in this study, we conclude that RF is much practical then ET in this specific problem.

5.2.2 Data Collection and dataset construction

The data was collected based on screening previously published papers involving recording the efficiency of the PSC devices, while providing the detail of the materials and the proprieties for each layer of the PSC, as well as illustrating the detail of the materials and methods used in the manufacturing process of these devices. The collection of all this information aims to determine -through ML- the materials, proprieties, and methods that have significant effects on the device PCE, where in this study we will focus on the effective factors. The information gathered can be classified into five categories:

-protection layer: which contains the subtract materials, and the encapsulation.

-ETL features: contains the thickness, materials, additive (doping), the deposition method and solvent, and the annealing temperature.

-Perovskite active layer features: contains the layer thickness, Materials short form (ABX), composition a ions, compositions b ions, composition x ions, the perovskite band gap, the additive materials, annealing temperature, deposition method, and solvent, deposition anti-solvent.

-HTL features: contains the layer thickness, layer materials, additives, deposition method, and the deposition solvent.

-Back contact features: contains the BC thickness, the BC materials, the deposition method, and the deposition solvent.

The data was gathered from different teams, papers, and reviews [51, 59, 76] while respecting several guidelines, for example, the data that contains the PCE test under high temperature or intense light were neglected, since these two factors influence the performance of the PSC device [121]. Moreover, the data was organized and labeled under the features listed above in the form of a matrix (table) to prepare it for further preprocessing stage, Table 4.1 illustrates the organized data. Each column represents the variables of a specific feature, for example, the ETL thickness, HTL annealing temperature, etc. while the rows represent a data sample. The total data samples are 3000 data points and the total features in this dataset is 35, as we have seen in the previous chapters, the high dimensionality is bad for the ML model. However, there are some samples that have lack in feature values. For instance, from 3000 samples there are 898 data sample that doesn't contain the perovskite band gap value. Hence, In order to keep these samples, we have adopted two approaches depending on the type of the missing variables. In the categorical feature (e.g. the features that contain the formula of the material), the missing values were replaced by a specific arbitrary category in each feature. In the numerical features like the band gap, we adopted a quite complicated approach which can be summarized as follows: at first, we separated the categorical features from the numerical features, then made a correlation matrix between the numerical dataset with the PCE target, which computed a pairwise correlation of features and excluding missing values based on the Pearson correlation coefficient (PCC) through the following equation:

Given the dataset $D = \{(x_1, y_1), \ldots,(x_i, y_i)\}$ where $i \in \{1, \ldots., N\}$

$$\rho_{xy} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x}) \sum_{i=1}^{N}(y_i - \bar{y})}} \qquad (13)$$

N is the number of samples, and $\bar{x}, \bar{y}$ are the mean values of the feature and target respectively. The results of the correlation matrix are illustrated in Figure 5.1. The features that have a low correlation coefficient are deleted and the features that show a relatively important correlation with the PCE have emerged with the dataset that contains the categorical features. The numerical features dropped are the perovskite thermal annealing temperature, the perovskite layer thickness, and the HTL layer thickness. And the features that relatively have a significant effect on the device PCE are: the ELT layer thickness, the ETL annealing temperature, the perovskite band gap, and the back contact layer thickness.

Table 4.1. Data of Perovskite Solar cells used for Machine Learning training process.

Features (35)	**ETL Materials**	**ETL Thickness (nm)**	**...**	**Perovskite**	**Bandgap (eV)**	**PCE (%)**
Cells						
Cell 1	TiO_2	40	...	FAMAPbBrI	1.73	12.1
Cell 2	SnO_2	80	...	MAPbI	1.61	8
Cell 3	PCBM60+BCP	220	...	MAPbI	1.61	13.06
....	……..	…....	...	...	…….	….
Cell 2998	PCBM-60	120	...	CsAgBiBr	2.39	5.5
Cell 2999	TiO_2	70	...	MAPbI	1.61	17
Cell 3000	C60+BCP	80	...	CsPbBrI	2.07	9.6

However, the remained numerical features still contain missing values. To solve this problem we have predicted the missing values using RF Regressor through the following strategy: first transform one of the features that contains missing values to a target and the rest of the features

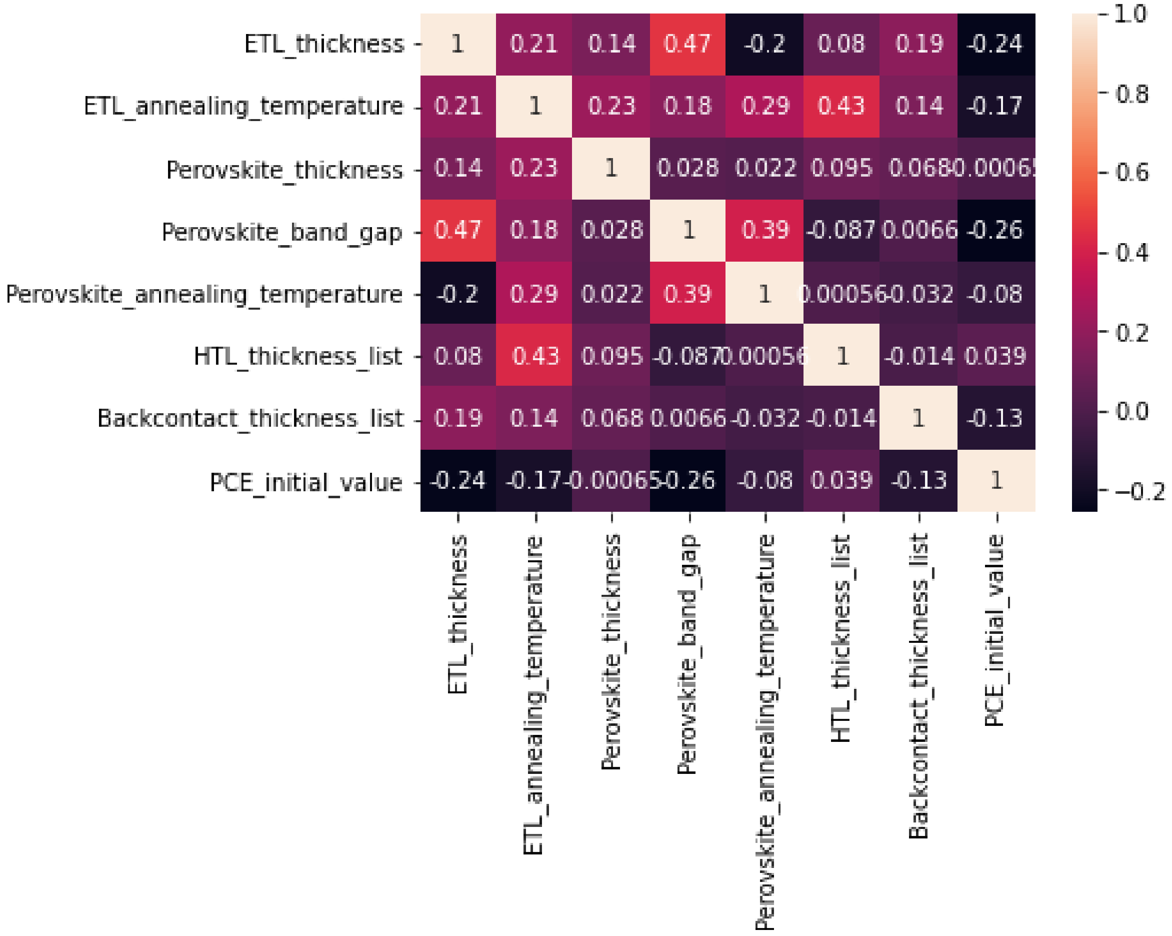


Figure 5.1. Correlation matrix of numerical features

and the PCE are considered as input for the ML model. Then, delete all the rows that contain missing values of this target. However, the numerical features still contain missing values. Hence, we have solved this problem by calculating the mean value of the feature column and replacing the missing values with it (as we have seen in section 3.3.2 this method can help to handle the numerical missing values). The resultant dataset was divided into 80% of the data to train the ML model and the rest for evaluating this model. If the accuracy of the resulting model is acceptable then we predicted the missing values of that target using the input from the samples that we have already deleted. Finally, we restored the missing value of the features that we had replaced with the mean value and repeated this process with every feature that contained missing values until all the missing values were replaced with predicted values. Overall, we have built five different RF models. These model's evaluation and the number of predicted values are summarized in Table 4.2.

5.2.3 Data Preprocessing

The preprocessing is a crucial step in preparing the dataset and making it appropriate for the ML model. Initially, we explored every data sample manually to ascertain the correctness of the data for both typing and organization. Then, we have encoded the categorical variables into numerical values. We have used GrisSearchCV for tuning the hyperparameters of the RF Regressor model, the results obtained are: the number of the trees in the forest is 200, the split criterion is "absolute error", and the maximum depth of the tree is 50. In particular, this dataset contains 32 different features that impact the ML model negatively due to the dimensionality curse (see Chapter 2).

Table 4.2- Illustration of the number of the predicted values and the accuracy of the model used.

Features	N° experimental values	N° predicted values	Accuracy of ML predictions
Bandgap	2102	898	87.9%
ETL thickness	1268	1738	96%
ETL annealing temperature	232	2768	98.5%
BMC thickness	2472	528	98%

Therefore, we have used the feature importance algorithm from the RF Regressor model to determine the relevant features to the PCE and delete the irrelevant ones. From 32 features, 17 feature was deleted and only 15 feature have relatively a significant importance score. The importance score of the relevant features is shown in Figure 5.2. By using the R^2 method the accuracy of the RF model resulted in 73.2% before feature selection (the ideal value is 100%). After the features selection process the accuracy of the model increased to 86.4%, and 1.3 by using mean square error (the optimum value is 0). The whole model building process is summarized in Figure 5.3.

5.2.4 Machine learning approach

The goal of this study is the attempt to optimize the materials and the proprieties of the PSC to increase the device PCE. We have determined the factors that have a significant effect on the device PCE through the feature importance technique. This means that improving these factors may enhance the PCE. In particular, these factors can be divided into two categories: numerical factors (ETL layer thickness, ETL deposition thermal annealing temperature, perovskite band gap, back contact thickness), and materials factors (ETL, perovskite, HTL, BC materials, deposition solvent, and anti-solvent). We have optimized the values of the numerical factors by generating three different PSC configurations as follows: Fe_2O_3/$CsPbBrI_2$/NiO-mp/Carbon, CdS/$FAMAPbI_3$/NiO-C/Au and PCBM-60/Phen-NaDPO/$MAPbI_3$/asy-PBTBDT/Ag (new configurations that didn't exist in the training dataset to prevent overfitting). Then, select one specific numerical factor continuously vary its value, and predict the device PCE every time the value changes. If there is a consistency of the variation of the PCE between the three configurations we generalize the optimum values. The results of this process are shown in Figures 5.4, 5.5, 5.6, 6.7. For optimizing the materials factors, we have followed a similar approach. We have taken the three PSC configurations and predicted the PCE for every material in the dataset, then, the materials that frequently appear in the three configurations that have the highest PCE are considered the more convenient. The results of this process are shown in Table 4.3.

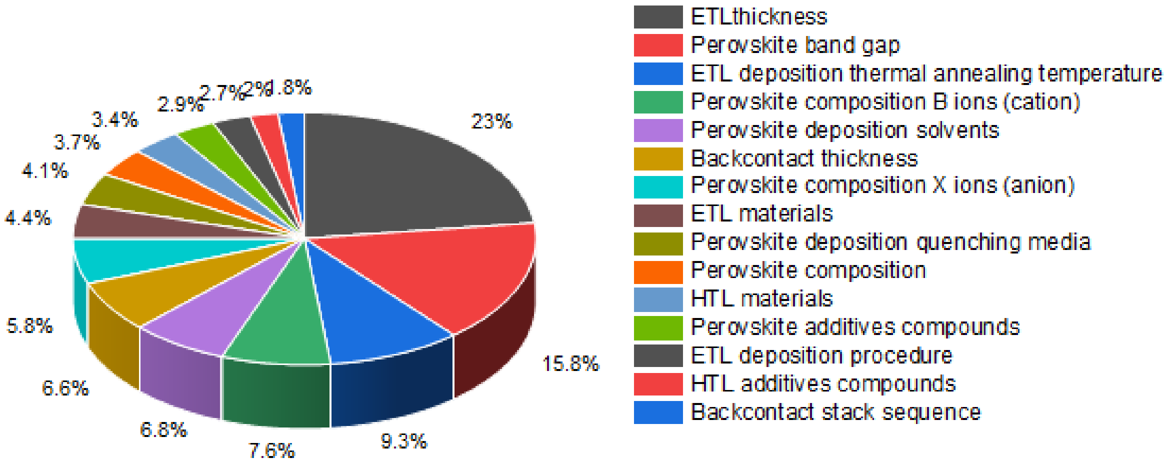


Figure 5.2- Importance score of the features that are relevant to the PCE

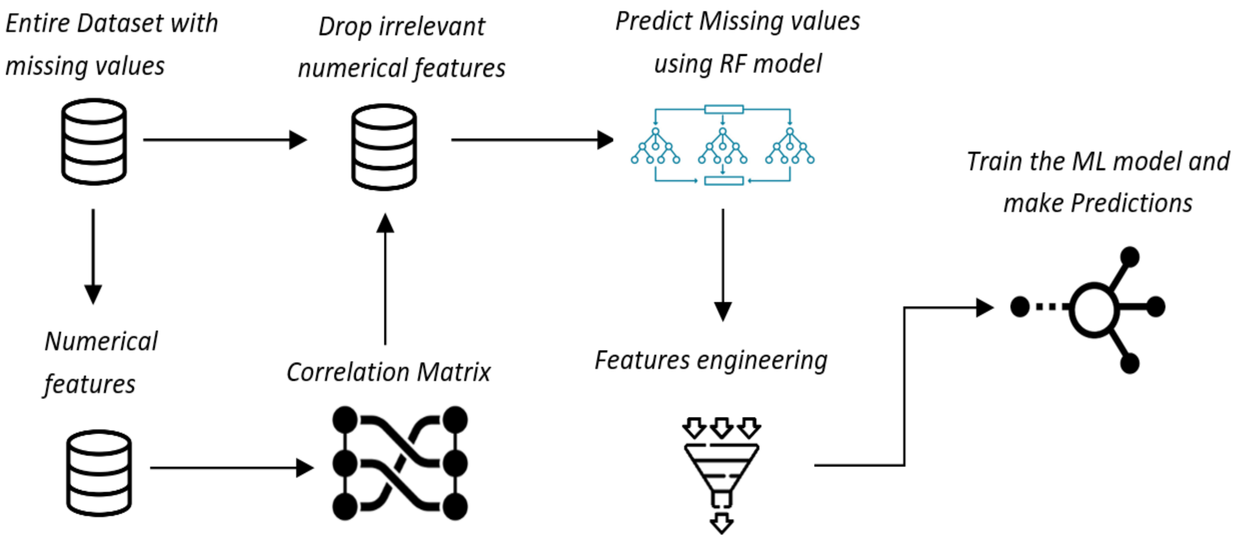

Entire Dataset with missing values
Drop irrelevant numerical features
Predict Missing values using RF model
Train the ML model and make Predictions
Numerical features
Correlation Matrix
Features engineering

Figure 5.3- Machine learning model building process workflow

5.3 Results and Discussion

5.3.1 Perovskite Material Band Gap

The band gap determines the light wavelength domain that can be absorbed by the active layer to convert it into electrical energy. So, it is a fundamental property of the photovoltaic semiconductors. A large band gap causes an increment of the Open Circuit Voltage V_{oc} which reduces the fill factor FF, hence reducing the device PCE. Otherwise, a low band gap decreases the amount of the photon absorption. Therefore, in order to balance the compromise between J_{sc} and VOC, the optimum band gap must be found.

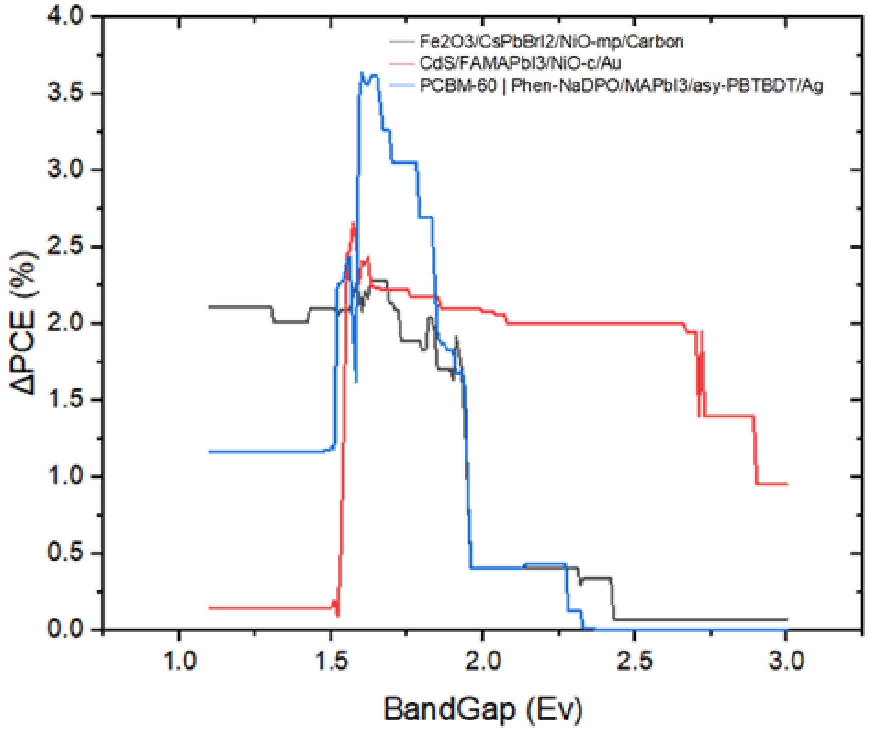


Figure 5.4- Variation of the PSC PCE in function of the perovskite material band gap.

Figure 5.4 shows the variation of the PSC PCE predicted using the RF Regressor algorithm. The band gap for the three configurations changes from 1.10 eV to 3.0 eV, with an increase of

0.01 eV for each prediction. The results show that the common optimum band gap values shared by the three configurations are between 1.55 eV and 1.60 eV. Furthermore, many experimental results show an evenly matched band gap range. For instance, A. Mahmud et all recorded 22.77% of PCE with a PSC device containing CsFAMAPbBrI as an active layer with a 1.6 eV band gap [122]. While J. Yoo et all reach 22.6% of PCE with a 1.56 eV band gap of FAMAPbBrI | (C6H13NH3)PbI perovskite material [123]. Interestingly, figure 5.2 shows that the band gap owes 15.8% of the importance score, which represents the second highest score among all the different features (35 features) indicating how important the band gap is for optimizing the device PCE.

5.3.2 ETL layer thickness

The charge transmission and the hole blocking provided by the Electron transport layer (ETL) plays a decisive role in the operation of the PSC devices. Moreover, from Figure 5.2. The total importance score of the factors related to the ETL layer is 39.4%, which makes it the second important layer after the perovskite active layer with a 46.5% total importance score. These two layers manage 85.9% of the device PCE according to the ML model. Which is considered a very interesting result. However, the ETL thickness is by far the most important factor for the PCE owing 23% of the importance score.

Figure 5.5 shows the effect of the variation of the ETL thickness on the device PCE for three different PSC structures. The variation of the ETL thickness value is from 20nm to 120000nm by adding 10nm for each prediction. The results show that the highest PCE increment occurs when the ETL thickness is between 140nm and 170nm. In particular, S. Sakib et all studied the ETL thickness effect on the PCE by simulating a PSC device using three different ETL materials (SnO_2, TiO_2, ZnO), they found that increasing the ETL thickness up to 200nm reduced the device PCE, and the optimum material found is SnO_2 [124]. Furthermore. Increasing ETL by more than 200nm causes a limitation in electron transit and raises the recombination rate. It also reduces the value of V_{oc} which in turn reduces the FF and the PCE.

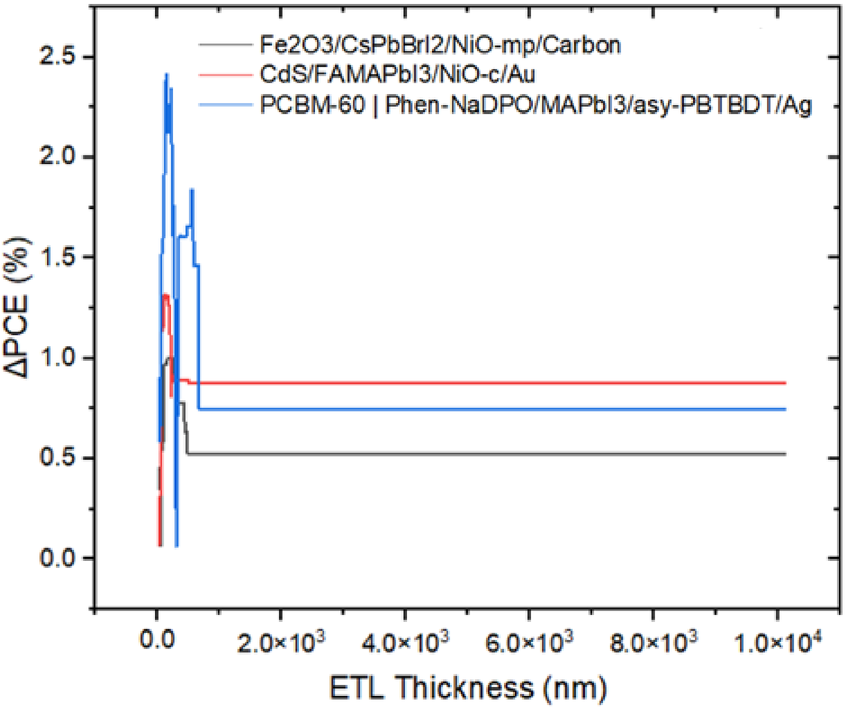


Figure 5.5- Variation of the PSC *PCE* in function of ETL material thickness.

5.3.3 ETL thermal annealing temperature

The thermal annealing temperature is an important factor in the material deposition phase during the device manufacturing process. In particular, the annealing temperature affects the electrical and optical properties of the materials deposited. It also can change the morphology of the films and influence the crystalline structure of the materials [125]. For instance. Y. Li et all show the influence of different annealing temperatures on the ETL layer consisting of ZnO film processed with the sol-gel technique. They found that the low annealing temperature results a smoother surface of the ZnO compared to the high annealing temperature. However, the low temperature results the highest PCE of 3.66%. Then, the PCE performance decreases with increasing annealing temperature, then intriguingly afterward starts to improve [126]. Similarly, the results in Figure 5.6 show an evenly matched PCE variation with an increased annealing temperature. In our study, The ETL annealing temperature range was taken between 25 ^{o}C and 550 ^{o}C with increments of 5 ^{o}C for each prediction. The results show that there is no coherence between the variations of PCE of the three PSC configurations. Hence, these results cannot be generalized, from which we conclude that each material has its own optimum thermal annealing temperature. This finding seems understandable since every material has its characteristic tolerance. Moreover, figure 5.2 shows that the thermal annealing temperature has a relatively significant impact on the device PCE with an importance score of 9.3%. Which indicates how important to select an adequate temperature.

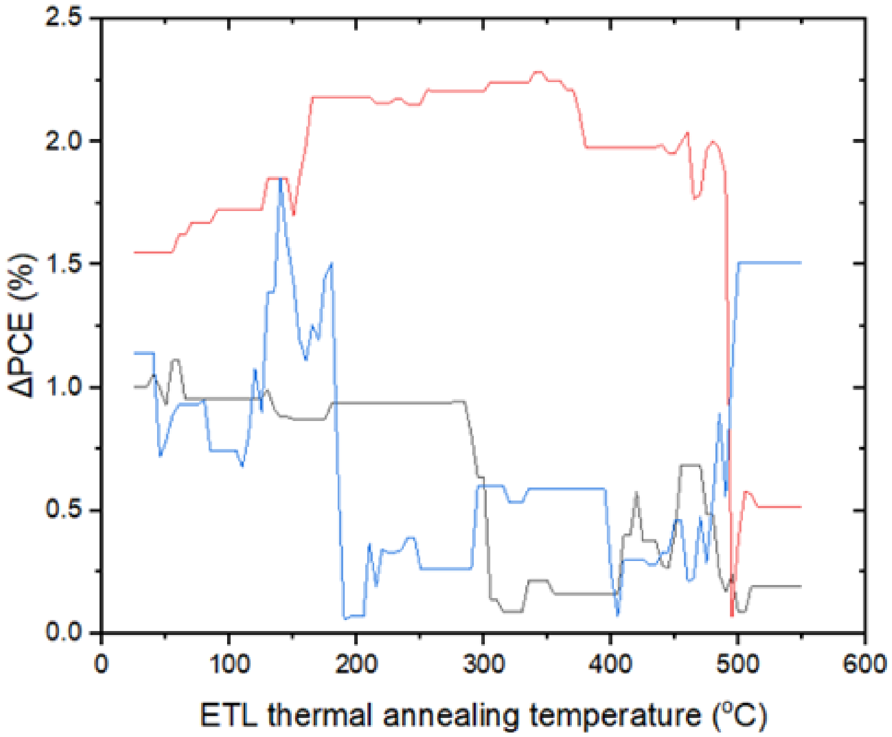


Figure 5.6- *PCE* variation as a function of ETL thermal annealing temperature

5.3.4 Back Contact thickness

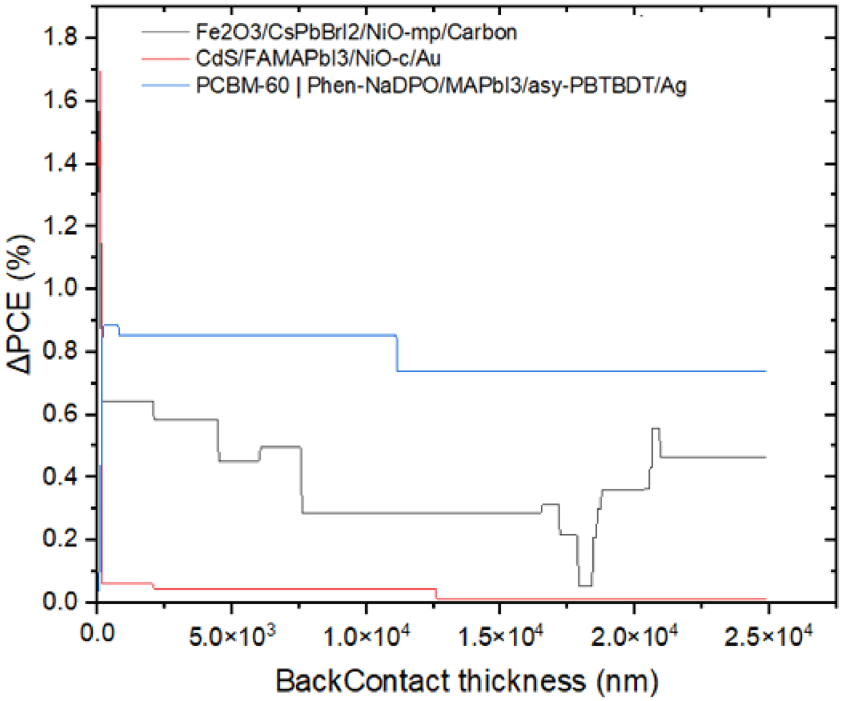


Figure 5.7- Variation of the PSC *PCE* in function of BMC thickness.

The Back contact electrode is an important layer for the PSC devices which prevents the absorption of the parasitic light brought by the transparent conductive electrodes and provides a potent feasibility to enhance the device PCE. Figure 5.2 shows that the back contact thickness has a 6.6% of importance score. Interestingly, this score is bigger than the back contact material score, it is apparent that the layer thickness of both back contact and ETL is more important than the materials used in this layer according to our ML model. Note that the layers of materials also have a significant importance. Figure 5.7 shows the variation of the device PCE with different back contact thicknesses. The thickness value varies from 1nm to 25000nm, with an increment of 50nm for each prediction. The results show that for every PSC structure, the value of the PCE strongly declines after 150nm of back contact thickness. Moreover, the optimum thickness was found to be less than 50nm. By using a computational method, A. Kang et all formed a PSC device with graphene back contact and found that the device PCE decreases when the back contact thickness increases from 7nm to 30nm and then remains constant after 30nm [127]. This is due to a rise in the lateral resistance of the back contact resulting from the increment of the number of graphene which reduces the fill factor.

5.3.5 Materials optimization

Figure 5.2 shows that the materials of the different PSC layers owing 18.2% of the importance score. In particular, the metal cation B composition of the perovskite (ABX_3) materials shows a significant relation with the device PCE with 7.6% of importance score. The most frequently used material as perovskite B composition is lead (Pb), and it has been used for almost all competent PSC [128]. However, due to the toxic nature and environmental harm of Pb, many researchers shifted to using lead-free or divalent mixed cation perovskite to reduce the Pb effect. For example, replacing the Pb with Tin (Sn) [129, 130]. Moreover, many research papers and reviews state that the B site has a significant impact on the device's stability and performance [128, 131].

Table 4.3 illustrates the materials from different layers that frequently appear in top-efficiency devices predicted by the RF Regressor model. The Total Materials column represents the number of different materials and compositions predicted for each layer. For example, in the ETL Layer, we investigate 262 different materials and material compositions (e.g. TiO_2 + PCBM). Furthermore, the SnO_2, PCBM, BCP, C60, and TiO_2 are respectively the most frequent materials in the top devices ETL layer. Hence, the optimum ETL material could be one of these materials or a composition between them, also it could be a mix with other different low-appearance materials. J. Kim et all prove that SnO_2 ETL has an excellent electron extraction and exceeds the widely used

TiO_2, which enhance rapidly the PCE, also it shows a higher stability over TiO_2 [132]. Dkhili et all found that the double layer ZnO and SnO_2 gives a higher PCE compared to SnO_2 only [133].

The optimum perovskite materials according to our ML model are CsFAMAPbBrI, MAPbI, and FAMAPbBrI respectively. Where the highest PCE was predicted owing to PSC with FAMAPbBrI as a perovskite active layer. L. Wang et all obtain a PCE of 22.7% by preparing a triple-cation mixed PSC with CsFAMAPbBrI as an active layer and a poly(triaryl amine) (PTAA) filled with spiro- OMeTAD, the ETL material used is SnO_2, and the back contact material is Au, the device also shows good stability which preserve 90% of its initial PCE for a 1000h under 85 ^{o}C temperature [134]. Where W. Wu et all reach 22.6% of PCE with a device composed of a modified $MAPbI_3$ as a perovskite active layer with a configuration of ITO/PTAA/$MAPbI_3$/C60|BCP/Cu [135].

The HTL optimum materials are spiro-OMeTAD with 20% of the samples in the highest PCE devices, then NiO and PTAA with 10%. However, spiro-OMeTAD remains by far the most frequently material used as HTL, and it is the first solid-state hole transport material used in PSCs. Which gives an efficient hole extraction and transmits to the back contact [136]. However, spiro-OMeTAD has a significant sensitivity to ion diffusion which causes stability degradation, and it also reduces the device PCE under thermal stress [137]. Moreover, Doping spiro-OMeTAD was found to enhance the device's stability and performance. For instance, found that doping spiro-OMeTad with trityltetra(pentafluoropheny)borate (TPP) improves the device stability and conductivity [138].

The optimum back contact material found is Au with a 30% appearance in the composition of the top device back contact followed by Ag with 26%. In particular. F. Behrouznejad et all compared a study with different back contact materials in a PSC configuration of FTO/m-

TiO_2/$MAPbI_3$/spiro-OMeTAD/Back contact, the material used are Au, Ag, Pt, Ni, Cu, Cr where performing a PCE of 16.4%, 16.5%, 14.7%, 7.8%, 9.2%, 0.04% respectively. Moreover, the PCE of Ag reduced quickly due to its instability which leave Au as the most adequate material as back contact [70]. Finally, the optimum deposition solvent and anti-solvent are DMF + DMSO + (other) and Chlorobenzene respectively.

Table 4.3- Different layer/deposition Materials that appear frequently in top PCE cells by using ML techniques

Layers/deposition materials	**Materials**	**Percentage**	**Total materials**
ETL	SnO_2	33 %	262
	PCBM-60	27 %	
	BCP	25 %	
	C60	23 %	
	TiO_2	23 %	
Perovskite	CsFAMAPbBrI	22 %	120
	MAPbI	15 %	
	FAMAPbBrI	11 %	
HTL	Spiro-MeOTAD	20 %	385
	NiOx	10 %	
	PTAA	10 %	
Back Contact	Au	30 %	65
	Ag	26 %	
	Cu	26 %	
	MoO_3	21 %	
	Ti	17 %	
Deposition Materials	DMF + DMSO (+other)	60 %	67
	DMSO + GBL		
Deposition quenching media	Chlorobenzee	58 %	30
	Ether	30 %	
	Toluene	29 %	

***Materials**: the material component of this layer in high-efficiency PSCs.*

***Percentage**: the number of current materials in top efficiency cells divided by the total materials in top cells.*
***Total materials**: the number of different materials compositions predicted by machine learning.*

5.4 Conclusion

In this chapter, we have followed a different approach aiming to optimize the PCE of the PSC device by using the ML technique of Random Forest Regressor and importance score. The RF algorithm was trained and evaluated by using 3000 samples of PSC experimental data from previous works. The relevant factors to the device PCE were investigated using the RF Importance score, where the materials were optimized by predicting the PCE of every material that exists in our dataset. The optimum ETL, back contact layer thickness was covered by predicting the PCE of three different PSC configurations with a wide range of values. As well as the optimum perovskite band gap.

General Conclusion

Summary

We have presented in this thesis a global investigation and analysis of all aspects of data related to the perovskite solar cells through the employment of advanced Machine learning techniques. The main methodological contribution of this research is the systematic investigation of the relevant factors that assist in achieving two main objectives: (i) enhancing the PSC device stability. (ii) Increasing the PSC device power conversion efficiency. The main difference between this approach and the previous approach is: the adoption in this research is entirely on the experimental data rather than using computational data. Even though the process of gathering experimental data is time-consuming and very hard, we thought that using experimental data is the most convenient approach for offering a better guide for future experimental research. The second advantage is that we have taken all the factors related to the PSC manufacturing and storage conditions. First, we provided an introduction to the application of different techniques including ML for material design and motivation about the latest achievement of the PSC devices. In Chapter 1, we have described machine learning while giving the working method of the algorithms used in this research, as well as we have defined some common technical terms used in this field. In Chapter 2, a comparison between three ML techniques concluded that the most appropriate algorithms for the material data are those that can separate non-linear data. For example, the neural networks, and the ensemble learning based on decision trees. In Chapters 3 and 4, the main factors that influence the PSC device stability degradation were analyzed in detail, where we found that the environmental conditions -especially the relative humidity- are key factors in the degradation of the PSC device. Moreover, we found that the hydrophobic materials are an adequate choice for mitigating this problem. In Chapter 5, the analysis of the prediction of PSC PCE results of ML involving a large scale of different materials and proprieties has helped us to determine the optimum factors for increasing the device PCE.

Future Outlook

There are significant opportunities for further research on using machine learning techniques for another type of material data. In the future, we intend to investigate the Optoelectronic devices upon the availability of the data.

References

[1] Hannah Ritchie, Max Roser and Pablo Rosado (2022) - "Energy". Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/energy' [Online Resource]

[2] Agrawal, A. & Choudhary, A. Perspective: Materials informatics and big data:

Realization of the “fourth paradigm” of science in materials science. Apl Materials 4, 053208 (2016)

[3] Chu, J. (no date) Moving past trial and error, MIT News | Massachusetts Institute of Technology. Available at: https://news.mit.edu/2012/profile-braatz-0215 (Accessed: 05 August 2023).

[4] Edison’s Lightbulb https://www.fi.edu/history- resources/edisonslightbulb. Accessed: 2020-06-16

[5] A. Jain*, S. P. O., G. Hautier, W. Chen, W.D. Richards, S. Dacek, S. Cholia, D. Gunter, D.

Skinner, G. Ceder, K.A. Persson, The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation. APL Mater. 2013, 1 (1), 011002

[6] White, A. The materials genome initiative: One year on. MRS Bulletin 37, 715–716 (2012).

[7] Kirklin, S., Saal, J., Meredig, B. *et al.* The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. *npj Comput Mater* **1**, 15010 (2015). https://doi.org/10.1038/npjcompumats.2015.10

[8] C., The Open Quantum Materials Database (Oqmd): Assessing the Accuracy of Dft Formation

Energies. npj Comput. Mater. 2015, 1 (1)

[9] Choudhary, K.; Garrity, K. F.; Reid, A. C.; DeCost, B.; Biacchi, A. J.; Walker, A. R. H.; Trautt, Z.; Hattrick-Simpers, J.; Kusne, A. G.; Centrone, A. J. a. p. a., The Joint Automated Repository for Various Integrated Simulations (Jarvis) for Data-Driven Materials Design. npj Comput Mater 2020, 6, 173.

[10] Choudhary, K.; DeCost, B.; Chen, C.; Jain, A.; Tavazza, F.; Cohn, R.; Park, C. W.; Choudhary, A.; Agrawal, A.; Billinge, S. J. L.; Holm, E.; Ong, S. P.; Wolverton, C., Recent Advances and Applications of Deep Learning Methods in Materials Science. npj Comput. Mater. 2022, 8 (1).

[11] Batra, R.; Song, L.; Ramprasad, R., Emerging Materials Intelligence Ecosystems Propelled by Machine Learning. Nat. Rev. Mater. 2020, 6 (8), 655.

[12] Moosavi, S. M.; Jablonka, K. M.; Smit, B., The Role of Machine Learning in the Understanding and Design of Materials. J Am Chem Soc 2020.

[13] Himanen, L.; Geurts, A.; Foster, A. S.; Rinke, P., Data-Driven Materials Science: Status, Challenges, and Perspectives. Adv Sci (Weinh) 2019, 6 (21), 1900808.

[14] Horton, M. K.; Dwaraknath, S.; Persson, K. A., Promises and Perils of Computational Materials Databases. Nat. Comput. Sci. 2021, 1 (1), 3

[15] Schleder, G.R. et al. (2019) ‘From DFT to machine learning: Recent approaches to materials science–A Review’, Journal of Physics: Materials, 2(3), p. 032001. doi:10.1088/2515-7639/ab084b.

[16] NationalCenter for Photovoltaics at the National NREL, Research cell efficiency records, Available online: https://www.nrel.gov/pv/cell-efficiency.html

[17] McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. The Bulletin of Mathematical Biophysics, 5(4), 115–133. doi:10.1007/bf02478259

[18] TURING, A. M. (1950). I.—COMPUTING MACHINERY AND INTELLIGENCE. Mind, LIX(236), 433–460. doi:10.1093/mind/lix.236.433

[19] Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers. IBM Journal of Research and Development, 3(3), 210–229. doi:10.1147/rd.33.0210

[20] Olivier Colliot. A non-technical introduction to machine learning. Olivier Colliot. Machine Learning for Brain Disorders, 197, Springer, 2023, Neuromethods. ffhal-03957125v3f

[21] [Waibel et al., 1989] Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., and Lang, K. (1989). Phoneme recognition using time-delay neural networks. Acoustics, Speech and Signal Processing, IEEE Transactions on, 37(3):328–339.

[22] Pomerleau, D. A. (1989). Alvinn: An autonomous land vehicle in a neural network. Technical report, DTIC Document.

[23] Campbell, M., Hoane, A. J., & Hsu, F. (2002). Deep Blue. Artificial Intelligence, 134(1-2), 57–83. doi:10.1016/s0004-3702(01)00129-1

[24] Ethem Alpaydin (2020). Introduction to Machine Learning (Fourth ed.). MIT. pp. xix, 1–3, 13–18. ISBN 978-0262043793.

[25] Kim, P. (2017). Matlab deep learning: With machine learning, neural networks and artificial intelligence. Apress. ISBN 978-1484228449

[26] Cristianini, N., Ricci, E. (2008). Support Vector Machines. In: Kao, MY. (eds) Encyclopedia of Algorithms. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-30162-4_415

[27] Aizerman, Mark A.; Braverman, Emmanuel M. & Rozonoer, Lev I. (1964). "Theoretical foundations of the potential function method in pattern recognition learning". Automation and Remote Control. 25: 821–837.

[28] Jin, Chi; Wang, Liwei (2012). Dimensionality dependent PAC-Bayes margin bound. Advances in Neural Information Processing Systems. CiteSeerX 10.1.1.420.3487. Archived from the original on 2015-04-02.

[29] Li, F., Peng, X., Wang, Z., Zhou, Y., Wu, Y., Jiang, M., & Xu, M. (2019). Machine Learning (ML)-Assisted Design and Fabrication for Solar Cells. ENERGY & ENVIRONMENTAL MATERIALS. doi:10.1002/eem2.12049

[30] Yao, W. *et al.* (2018) 'A support vector machine approach to estimate global solar radiation with the influence of fog and haze', *Renewable Energy*, 128, pp. 155–162. doi:10.1016/j.renene.2018.05.069.

[31] Belson, W. A. (1959). Matching and Prediction on the Principle of Biological Classification. Journal of the Royal Statistical Society. Series C (Applied Statistics), 8(2), 65–75. https://doi.org/10.2307/2985543

[32] Breiman, L. (2001). Random forests. Machine Learning, 45(1):5–32.

[33] Schmidhuber, Jürgen (2015-01-01). "Deep learning in neural networks: An overview". Neural Networks. 61: 85–117. arXiv:1404.7828. doi:10.1016/j.neunet.2014.09.003. ISSN 0893-6080. PMID 25462637. S2CID 11715509.

[34] D. Y. Singh and A. S. Chauhan, "Neural networks in data mining," Journal of Theoretical and Applied Information Technology, vol. 5, no. 1, pp. 37–42, 2009.

[35]]Tian Xie,T, X,(2020),Deep Learning Methods for the Design and

Understanding of Solid Materials[Doctoral thesis,MASSACHUSETTS INSTITUTE OF TECHNOLOGY

[36] Gómez-Bombarelli, R. et al. Design of efficient molecular organic light-emitting

diodes by a high-throughput virtual screening and experimental approach. Nature materials 15, 1120–1127 (2016).

[37] Fujimura, K. et al. Accelerated Materials Design of Lithium Superionic Conductors Based on First-Principles Calculations and Machine Learning Algorithms.

Advanced Energy Materials 3, 980–985 (2013).

[38]]Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine

learning for molecular and materials science. Nature 559, 547–555 (2018).

[39] Schmidt, J., Marques, M. R., Botti, S. & Marques, M. A. Recent advances and

applications of machine learning in solid-state materials science. npj Computational Materials 5, 1–36 (2019)

[40] Vandermause, J., Torrisi, S. B., Batzner, S., Kolpak, A. M. & Kozinsky, B. On-the-fly Bayesian active learning of interpretable force-fields for atomistic rare events. arXiv preprint arXiv:1904.02042 (2019

[41] Engel, E. A., Anelli, A., Ceriotti, M., Pickard, C. J. & Needs, R. J. Mapping uncharted territory in ice from zeolite networks to ice structures. Nature communications 9, 1–7 (2018)

[42] Pan, M., Li, C., Gao, R., Huang, Y., You, H., Gu, T., & Qin, F. (2020). Photovoltaic power forecasting based on a support vector machine with improved ant colony optimization. Journal of Cleaner Production, 123948. doi:10.1016/j.jclepro.2020.123948

[43] Khondoker, F., Rao, S., Spanias, A., & Tepedelenlioglu, C. (2018). Photovoltaic Array Simulation and Fault Prediction via Multilayer Perceptron Models. 2018 9th International Conference on Information, Intelligence, Systems and Applications (IISA). doi:10.1109/iisa.2018.8633699

[44] Vieira, R.G. et al. (2022) 'Comparing multilayer perceptron and probabilistic neural network for PV Systems Fault Detection', Expert Systems with Applications, 201, p. 117248. doi:10.1016/j.eswa.2022.117248.

[45] Sahoo, S. K., Manoharan, B., & Sivakumar, N. (2018). Introduction. Perovskite Photovoltaics, 1–24.

[46] Kojima A, Teshima K, Shirai Y, Miyasaka T: Organometal halide perovskites as visible-light sensitizers for photovoltaic cells, J Am Chem Soc 131:6050–6051, 2009.

[47] Im J-H, Lee C-R, Lee J-W, Park S-W, Park N-G: 6.5% efficient perovskite quantum-dot-sensitized solar cell, Nanoscale 3:4088–4093, 2011.

[48] Kim H-S, Lee C-R, Im J-H, Lee K-B, Moehl T, Marchioro A, et al: Lead iodide perovskite sensitized all-solid-state submicron thin film mesoscopic solar cell with efficiency exceeding 9%, Sci Rep 2:591, 2012

[49] Li, Y. et al. (2023) 'All-inorganic perovskite solar cells featuring mixed group Iva cations', Nanoscale, 15(16), pp. 7249–7260. doi:10.1039/d3nr00133d.

[50] Zhou, D. et al. (2018) ‘Perovskite-based solar cells: Materials, methods, and future perspectives’, Journal of Nanomaterials, 2018, pp. 1–15. doi:10.1155/2018/8148072.

[51] Jacobsson, T. J., Hultqvist, A., García-Fernández, A., Anand, A., Al-Ashouri, A., Hagfeldt, A., Crovetto, A., Abate, A., Ricciar-dulli, A. G., Vijayan, A., Kulkarni, A., Anderson, A. Y., Darwich, B. P., Yang, B., Coles, B. L., Perini, C. A. R., Rehermann, C., Ramirez, D., Fairen-Jimenez, D., ... Unger, E. (2022). An open-access database and analysis tool for perovskite solar cells based on the FAIR data principles. Nature Energy, 7(1), 107–115.doi.org/10.1038/s41560-021-00941-3.10

[52] Kusy, M. and Kowalski, P.A. (2018) ‘Weighted Probabilistic Neural Network’, Information Sciences, 430–431, pp. 65–76. doi:10.1016/j.ins.2017.11.036.

[53] C. Silverman, D. (2010) TUTORIAL ON ARTIFICIAL NEURAL NETWORKS, argentumsolutions.com. Available at: https://web.archive.org/web/20101212042242/, http://argentumsolutions.com/tutorials/neural_tut rialpg8.html (Accessed: 20 July 2023).

[54] Pedregosa, F., Varoquaux, Gael, Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... others. (2011). Scikit-learn: Machine learning in Python.Journal of Machine Learning Research,12(Oct), 2825–2830.

[55] Abadi, Martin, Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... others. (2016). Tensorflow: A system for large-scalemachine learning. In12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)(pp. 265–283)

[56] Li, J., Pradhan, B., Gaur, S., & Thomas, J. (2019). Predictions and Strategies Learned from Machine Learning to Develop High-Performing Perovskite Solar Cells. Advanced Energy Materials, 9(46), 1901891.doi.org/10.1002/aenm.201901891.23.

[57] Anguita, D., Ghio, A., Greco, N., Oneto, L., & Ridella, S. (2010). Model selection for support vector machines: Advantages and disadvantages of the Machine Learning Theory. The 2010 International Joint Conference on Neural Networks (IJCNN), 1–8. doi.org/10.1109/IJCNN.2010.5596450.24

[58] Pal, S. K., & Mitra, S. (1992). Multilayer perceptron, fuzzy sets, and classification. IEEE Transactions on Neural Networks, 3(5), 683–697.doi.org/10.1109/72.159058.25

[59] Khalid Salama. (2021). Probabilistic Bayesian Neural Networks. Available Online: {https://keras.io/examples/keras_reci-pes/bayesian_neural_networks/}

[60] Ç. Odabaşı and R. Yıldırım, "Machine learning analysis on stability of perovskite solarcells,"SolarEnergyMaterialsandSolarCells,vol.205,Feb.2020,doi:10.1016/j.solmat.2019.110 284.

[61] G. Gordillo, O. G. Torres, M. C. Abella, J. C. Peña, and O. Virguez, "Improving the stability of MAPbI3 films by using a new synthesis route," Journal of Materials Research and Technology, vol.9,no.6, pp. 13759–13769,Nov.2020, doi:10.1016/j.jmrt.2020.09.095.

[62] Faber, F. A.; Lindmaa, A.; von Lilienfeld, O. A.; Armiento, R.Machine learning energiesof 2 million elpasolite (ABC2D6) crystals.Phys. Rev. Lett. 2016, 117, 135502.6.

[63] Meredig, B.; Agrawal, A.; Kirklin, S.; Saal, J. E.; Doak,J. W.;Thompson, A.; Zhang, K.; Choudhary, A.; Wolverton, C. Combinatorial screening for new materials in unconstrained composition spacewith machine learning. Phys. Rev. B: Condens. Matter Mater. Phys. 2014, 89, 094104.7.

[64] Oliynyk, A. O.; Antono, E.; Sparks, T. D.; Ghadbeigi, L.;Gaultois, M. W.; Meredig, B.; Mar, A. High-throughput machinelearning-driven synthesis of full-Heusler compounds. Chem. Mater.2016, 28, 7324−7331

[65] J. Schmidt, J. Shi, P. Borlido, L. Chen, S. Botti, and M. A. L. Marques, "Predicting theThermodynamic Stability of Solids Combining Density Functional Theory and MachineLearning,"ChemistryofMaterials,vol.29,no.12,pp.5090 5103, Jun.2017, doi: 10.1021/acs.chemmater. 7b00156

[66] B.Naman, How does ExtraTreesClassifier reduce the risk of overfitting, Oct.22,2018

[67] T. Chen and C. Guestrin, "XGBoost," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Aug. 2016, pp. 785–794.doi:10.1145/2939672.2939785

[68] Hastie, T., Friedman, J. and Tisbshirani, R. (2017) The elements of Statistical Learning: Data Mining, Inference, and prediction. New York: Springer.

[69] Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984) Classification and Regression Trees. Chapman and Hall, Wadsworth, New York.

[70] F. Behrouznejad, S. Shahbazi, N. Taghavinia, H. P. Wu, and E. Wei-GuangDiau, "A studyon utilizing different metals as the back contact of CH3NH3PbI3 perovskite solar cells,"Journal of Materials Chemistry A,vol.4,no.35,pp.13488–13498,2016,doi:10.1039/c6ta05938d.

[71] Farhadi, B., Ciprian, M., Zabihi, F., & Liu, A. (2021). Influence of contact electrode and light power on the efficiency of tandem perovskite solar cell: Numerical simulation. Solar Energy, 226, 161–172. doi:10.1016/j.solener.2021.08.043

[72] M. Shahbazi and H. Wang, “Progress in research on the stability of organometal perovskite solar cells,” Solar Energy, vol. 123. Elsevier Ltd, pp. 74–87, Jan. 01, 2016. doi:10.1016/j.solener.2015.11.008.

[73] P. Geurts, D. Ernst, L. Wehenkel, “Extremely randomized trees”. Mach Learn, vol. 63, pp. 3–42, 2006. doi:10.1007/s10994-006-6226-1.

[74] R. Wang, M. Mujahid, Y. Duan, Z.-K. Wang, J. Xue, Y. Yang, “A Review of Perovskites Solar Cell Stability”. Advanced Functional Materials, vol. 29, no. 47, 1808843, 2019, doi: 10.1002/adfm.201808843.

[75] B. Yılmaz, R. Yıldırım, “Critical review of machine learning applications in perovskite solar research”. Nano Energy, vol. 80, pp. 105546, 2021 doi:10.1016/j.nanoen.2020.105546, [76] Ç. Odabaşı-Özer, R. Yıldırım, “Performance analysis of perovskite solar cells in 2013–2018 using machine-learning tools”. Nano Energy, vol. 56, pp. 770-791 2018. doi:10.1016/j.nanoen.2018.11.069.

[76] Zhang, Z. et al. (2022) ‘Big data driven perovskite solar cell stability analysis’, Nature Communications, 13(1). doi:10.1038/s41467-022-35400-4.

[77] K. Kenji, A. R.Larry, “The Feature Selection Problem: Traditional Methods and a New Algorithm,” , AAAI, vol. 2, pp. 129-134, 1992

[78] M. Kuhn, K. Johnson, “An introduction to feature selection. In Applied predictive modelling”, pp. 487-519. Springer, New York, NY, 2013, doi:10.1007/978-1-4614-6849-3_19.

[79] J. Brank, M. Grobelnik, N. Milic-Frayling, D. Mladenic, “Feature Selection,” in Encyclopedia of Machine Learning, Boston, MA: Springer US, 2011, pp. 402–406. doi: 10.1007/978-0-387-30164-8_306.

[80] Kuhn, M. and Johnson, K. (2016) Applied predictive modeling.pp. 28 and 487-490 New York: Springer.

[81] Nina, Z. (2015, January 5). Random Test/Train Split is not Always Enough. Win Vector LLC Data Science Advising, Consulting, and Training

[82] Cheng, Y., & Ding, L. (2021). Pushing commercialization of perovskite solar cells by improving their intrinsic stability. Energy & Environmental Science, 14(6), 3233–3255. doi:10.1039/d1ee00493j

[83] X. Zeng, T. Zhou, C. Leng, Z. Zang, M. Wang, W. Hu, X. Tang, S. Lu, L. Fang, M. Zhou, "Performance improvement of perovskite solar cells by employing a CdSe quantum dot/PCBM composite as an electron transport layer", Journal of Materials Chemistry A, vol. 5, no. 33, pp. 17499-17505, 2017, doi: 10.1039/C7TA00203C

[84] F. K. Aldibaja, L. Badia, E. Mas-Marzá, R. S. Sánchez, E. M. Barea, I. Mora-Sero, "Effect of different lead precursors on perovskite solar cell performance and stability". Journal of Materials Chemistry A, vol. 3, no. 17, pp. 9194–9200, 2015. doi:10.1039/c4ta06198e.

[85] F. Arabpour Roghabadi, M. Alidaei, S. M. Mousavi, T. Ashjari, A. S. Tehrani, , V. Ahmadi, S. M. Sadrameli, "Stability progress of perovskite solar cells dependent on the crystalline structure: From 3D ABX3 to 2D Ruddlesden–Popper perovskite absorbers". Journal of Materials Chemistry A, vol. 7, no. 11, pp. 5898–5933. doi:10.1039/c8ta10444a.

[86] J. Zhao, X. Zheng, Y. Deng, T. Li, Y. Shao, A. Gruverman, J. Huang, "Is Cu a stable electrode material in hybrid perovskite solar cells for a 30-year lifetime?" Energy & Environmental Science, vol. 9, no. 12, pp. 3650–3656. doi:10.1039/c6ee02980a.

[87] Han, Y., Meyer, S., Dkhissi, Y., Weber, K., Pringle, J. M., Bach, U., … Cheng, Y.-B. "Degradation observations of encapsulated planar CH3NH3PbI3 perovskite solar cells at high temperatures and humidity". Journal of Materials Chemistry A, vol. 3, no. 15, pp. 8139–8147, 2015,doi:10.1039/c5ta00358j.

[88] B. Brunetti, C. Cavallo, A. Ciccioli, G. Gigli, A. Latini, "On the thermal and thermodynamic (In)Stability of methylammonium lead halide perovskites", Sci. Rep. vol. 6, pp. 31896, 2016. doi:10.1038/srep31896.

[89] J. H. Noh, S.H. Im, J.H. Heo, T.N. Mandal, S. Il Seok, "Chemical management for colorful, efficient, and stable inorganic-organic hybrid nanostructured solar cells", Nano Lett, vol. 13, pp. 1764–1769, 2013. doi:10.1021/nl400349b.

[90] K. O. Ogunniran, N. T. Martins, "Humidity and Moisture Degradation of Perovskite Material in Solar Cells: Effects on Efficiency". IOP Conference Series: Earth and Environmental Science, vol. 655, no. 1, 012049, 2021. doi:10.1088/1755-1315/655/1/012049.

[91] J. M. Frost, K. T. Butler, F. Brivio, C. H.Hendon, M. van Schilfgaarde, A. Walsh, "Atomistic Origins of High-Performance in Hybrid Halide Perovskite Solar Cells". Nano Letters, vol. 14, no. 5, pp. 2584–2590, 2014. doi:10.1021/nl500390f.

[92] S. Pont, D. Bryant, C.-T. Lin; N. Aristidou, S. Wheeler, X. Ma, R. Godin, S. A. Haque, J. R. Durrant, "Tuning CH3NH3Pb-(I1-xBrx)3 Perovskite Oxygen Stability in Thin Films and Solar Cells". J. Mater. Chem. A, vol. 5, pp. 9553−9560, 2017. doi: 10.1039/C7TA00058H.

[93] B. A. Nejand, V. Ahmadi, S. Gharibzadeh and H. R. Shahverdi, "Cuprous oxide as a potential low-cost hole-transport material for stable perovskite solar cells",ChemSusChem, vol. 9, pp. 302-313. 2016. doi: 10.1002/cssc.201501273.

[94] X. Liu, , Y. Zhang, J. Hua, Y. Peng, , F. Huang, , J. Zhong, , Y. Cheng, (2018). Improving the intrinsic thermal stability of the MAPbI3 perovskite by incorporating cesium 5-aminovaleric acetate. RSC Advances, vol. 8, no. 27, pp. 14991–14994. doi:10.1039/c7ra13611k

[95] N. Pellet, P. Gao, G. Gregori, T.-Y. Yang, M. K. Nazeeruddin, J. Maier, M. Grätzel, “Mixed-organic-cation perovskite photovoltaics for enhanced solar-light harvesting”, Angew. Chem. Int'l Ed, vol. 53, pp. 3151-3157, 2014. doi: 10.1002/ange.201309361.

[96] N. J. Jeon, J. H. Noh, W. S. Yang, Y. C. Kim, S. Ryu, J. Seo, S. I. Seok, “Compositional engineering of perovskite materials for high-performance solar cells”, Nature, vol. 517, pp. 476-480, 2015. doi: 10.1038/nature14133.

[97] E.-B. Kim, M. S. Akhtar, H.-S. Shin, S. Ameen, M. K. Nazeeruddin, “A review on two-dimensional (2D) and 2D-3D multidimensional perovskite solar cells: Perovskites structures, stability, and photovoltaic performances”. Journal of Photochemistry and Photobiology C: Photochemistry Reviews, vol. 48, pp. 100405. doi:10.1016/j.jphotochemrev.2021.100405

[98] C. Ma, C. Leng, Y. Ji, X. Wei, K. Sun, L. Tang, J. Yang, W. Lou, C. Li, Y. Deng, S. Feng, J. Shen, S. Lu, C. Du, H. Shi, “2D/3D perovskite hybrids as moisturetolerant and efficient light absorbers for solar cells”, Nanoscale, vol. 8, pp. 18309–18314, 2016. doi: 101039/C6NR04741F.

[99] G. Grancini, C. Roldan-Carmona, I. Zimmermann, E. Mosconi, X. Lee, D. Martineau, S. Narbey, F. Oswald, F. De Angelis, M. Graetzel, et al., “One-year stable perovskite solar cells by 2D/3D interface engineering”, Nat. Commun, vol. 8, pp. 15684, 2017. doi: 10.1038/ncomms15684.

[100] J. Yuan, Y. Jiang, T. He, G. Shi, Z. Fan, M. Yuan, “Two-dimensional perovskite capping layer for stable and efficient tin-lead perovskite solar cells”, Sci. China Chem, vol. 62, pp. 629–636, 2019. doi:10.1007/s11426-018-9436-1.

[101] C. Liu, Y. Yang, Y. Ding, J. Xu, X. Liu, B. Zhang, S. Dai, “High-Efficiency and UV-Stable Planar Perovskite Solar Cells Using a Low-Temperature, Solution-Processed Electron-Transport Layer”. ChemSusChem, vol. 11, no. 7, pp. 1232–1237, 2018. doi:10.1002/cssc.201702248.

[102] G. Niu, X. Guo, L. Wang, “Review of recent progress in chemical stability of perovskite solar cells”. Journal of Materials Chemistry A, vol. 3, no. 17, pp. 8970–8980. doi:10.1039/c4ta04994b.

[103] A. F ujishima, T. N. Rao, D. A. Tryk, “Titanium dioxide photocatalysis”. Journal of Photochemistry and Photobiology C: Photochemistry Reviews, vol. 1, no. 1, pp. 1–21. doi:10.1016/s1389-5567(00)00002-2

[104] S. Ito, S. Tanaka, K. Manabe, H. Nishino, “Effects of Surface Blocking Layer of Sb2S3 on Nanocrystalline TiO2 for CH3NH3PbI3 Perovskite Solar Cells”. The Journal of Physical Chemistry C, pp. 118, no. 30, pp. 16995–17000. doi:10.1021/jp500449z.

[105] S. K. Pathak, A. Abate, P. Ruckdeschel, B. Roose, K. C. Gödel, Y. Vaynzof, U. Steiner, “Performance and Stability Enhancement of Dye-Sensitized and Perovskite Solar Cells by Al Doping of TiO2”. Advanced Functional Materials, vol. 24, no. 38, pp. 6046–6055. doi:10.1002/adfm.201401658.

[106] G. Yin, J. Ma, H. Jiang, J. Li, D. Yang, F. Gao, S. F. Liu, “Enhancing Efficiency and Stability of Perovskite Solar Cells through Nb-Doping of TiO2 at Low Temperature”. ACS Applied Materials & Interfaces, vol. 9, no. 12, pp. 10752–10758. doi:10.1021/acsami.7b01063

[107] X. F. Wang, L. Wang, N. Tamai, O. Kitao, H. Tamiaki, S. I. Sasaki, “Development of Solar Cells Based on Synthetic NearInfrared Absorbing Purpurins: Observation of Multiple Electron Injection Pathways at Cyclic Tetrapyrrole-Semiconductor Interface”. J. Phys. Chem. C, vol. 115, no. 49, pp. 24394−24402, 2011. doi: 10.1021/jp206206x

[108] K. Junu, S. K. Kwang, W. M. Chang, “Efficient electron extraction of SnO2 electron transport layer for lead halide perovskite solar cell”, nature, vol. 100, 2020. doi: 10.1038/s41524-020-00370-y

[109] A. Mei, X. Li, L. Liu, Z. Ku, T. Liu, Y. Rong, H. Han, “A hole-conductor-free, fully printable mesoscopic perovskite solar cell with high stability”. Science, vol. 345, no. 6194, pp. 295–298, 2014. doi:10.1126/science.1254763.

[110] M. Spalla, L. Perrin, E. Planes, M. Matheron, S. Berson, L. Flandin, “Effect of the hole transporting / active layer interface on the perovskite solar cell stability”. ACS Applied Energy Materials, vol. 3, no. 4, pp. 3282-3292 2020. doi:10.1021/acsaem.9b02281

[111] N. Arora, M. I. Dar, M. Abdi-Jalebi, F. Giordano, N. Pe llet, G.Jacopin, M. Grätzel, “Intrinsic and Extrinsic Stability of Formamidinium Lead Bromide Perovskite Solar Cells Yielding High Photovoltage”. Nano Letters, vol. 16, no. 11, pp. 7155–7162, 2016. doi:10.1021/acs.nanolett.6b03455

[112] V. Trifiletti, V. Roiati , S. Clella, R. Giannuzzi, L. De Marco, A. Rizzo, A. Gigli, G, “NiO/MAPbI3-xClx/PCBM: A Model Case for an Improved Understanding of Inverted Mesoscopic Solar Cells”. ACS Applied Materials & Interfaces, vol. 7, no. 7, pp. 4283–4289, 2015. doi:10.1021/am508678p.

[113] E. M. Younes, A. Gurung, B. Bahrami, E. M. El-Maghraby, Q. Qiao, “Enhancing efficiency and stability of inverted structure perovskite solar cells with fullerene C60 doped PC61BM electron transport layer”. Carbon, vol. 180, pp. 226–236, 2021. doi:10.1016/j.carbon.2021.05.008

[114] Ye, J., Zheng, H., Zhu, L., Zhang, X., Jiang, L., Chen, W., … Dai, S. (2017). High-temperature shaping perovskite film crystallization for solar cell fast preparation. Solar Energy Materials and Solar Cells, 160, 60–66. doi:10.1016/j.solmat.2016.10.02

[115] Ye, J., Zhu, L., Zhou, L., Liu, X., Zhang, X., Zheng, H., … Dai, S. (2016). Effective and reproducible method for preparing low defects perovskite film toward highly photoelectric properties with large fill factor by shaping capping layer. Solar Energy, 136, 505–514. doi:10.1016/j.solener.2016.07.034

[116] Zhang, T., Meng, X., Bai, Y., Xiao, S., Hu, C., Yang, Y., … Yang, S. (2017). Profiling the organic cation-dependent degradation of organolead halide perovskite solar cells. Journal of Materials Chemistry A, 5(3), 1103–1111. doi:10.1039/c6ta09687e

[117] Shalan, A. E., Oshikiri, T., Narra, S., Elshanawany, M. M., Ueno, K., Wu, H.-P., … Misawa, H. (2016). Cobalt Oxide (CoOx) as an Efficient Hole-Extracting Layer for High-Performance Inverted Planar Perovskite Solar Cells. ACS Applied Materials & Interfaces, 8(49), 33592–33600. doi:10.1021/acsami.6b10803

[118] T. Wu, J. Wang, “Deep Mining Stable and Nontoxic Hybrid Organic–Inorganic Perovskites for Photovoltaics via Progressive Machine Learning”. ACS Applied Materials & Interfaces, vol. 12, no. 52, pp. 57821-57831, 2020. doi:10.1021/acsami.0c10371

[119] Ç. Odabaşı, R. Yıldırım, “Assessment of Reproducibility, Hysteresis and Stability Relations in Perovskite Solar Cells Using Machine Learning. Energy Technology”, vol. 8, no. 12, pp. 1901449, doi:10.1002/ente.201901449.

[120] Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

[121] Pilar Lopez-Varo, Mohamed Amara, Stefania Cacovich, Arthur Julien, Armelle Yaiche, et al.. Dynamic temperature effects in perovskite solar cells and energy yield. Sustainable Energy & Fuels, 2021, 5 (21), pp.5523-5534. ff10.1039/d1se01381eff. ffhal-03622152f

[122] Mahmud, M.A. et al. (2021) ‘Origin of efficiency and stability enhancement in high-performing mixed dimensional 2D-3D perovskite solar cells: A Review’, Advanced Functional Materials, 32(3), p. 2009164. doi:10.1002/adfm.202009164.

[123] Yoo, J.J. et al. (2019) ‘An interface stabilized perovskite solar cell with high stabilized efficiency and low voltage loss’, Energy &amp;amp; Environmental Science, 12(7), pp. 2192–2199. doi:10.1039/c9ee00751b.

[124] Sakib, S. et al. (2023) ‘Effect of transport layer thickness in lead-based perovskite solar cell: A numerical simulation’, Materials Today: Proceedings, 80, pp. 1022–1026. doi:10.1016/j.matpr.2022.11.456.

[125] Timofeev, V.A. et al. (2020) ‘Effect of annealing temperature on the morphology, structure, and optical properties of nanostructured sno(x) films’, Materials Research Express, 7(1), p. 015027. doi:10.1088/2053-1591/ab6122.

[126] Li, Y. et al. (2015) ‘Influence of annealing temperature of zno film as the electron transport layer on the performance of Polymer Solar Cells’, Optoelectronics Letters, 11(4), pp. 260–263. doi:10.1007/s11801-015-5072-4.

[127] Kang, A.K., Zandi, M.H. and Gorji, N.E. (2019) ‘Simulation analysis of graphene contacted perovskite solar cells using SCAPS-1D’, Optical and Quantum Electronics, 51(4). doi:10.1007/s11082-019-1802-3.

[128] Khalid, M. and Mallick, T.K. (2023) ‘Stability and performance enhancement of Perovskite Solar Cells: A Review’, Energies, 16(10), p. 4031. doi:10.3390/en16104031.

[129] Soykan, C.; Gocmez, H. The Physical Properties of Bismuth Replacement in Lead Halogen Perovskite Solar Cells: CH3NH3Pb1−xBixI3 Compounds by Ab-Initio Calculations. Results Phys. 2019, 13, 102278. [Google Scholar] [CrossRef]

[130] Zhang, Q.; Hao, F.; Li, J.; Zhou, Y.; Wei, Y.; Lin, H. Perovskite Solar Cells: Must Lead Be Replaced–and Can It Be Done? Sci. Technol. Adv. Mater. 2018, 19, 425–442.

[131] Nyayban, A., Panda, S. and Chowdhury, A. (2023) ‘The effect of B-site alloying on the electronic and opto-electronic properties of rbpbi3: A DFT study’, Physica B: Condensed Matter, 649, p. 414384. doi:10.1016/j.physb.2022.414384.

[132] Kim, J., Kim, K.S. and Myung, C.W. (2020) ‘Efficient electron extraction of SNO2 electron transport layer for lead halide perovskite solar cell’, npj Computational Materials, 6(1). doi:10.1038/s41524-020-00370-y.

[133] Dkhili, M. et al. (2022) ‘Attributes of high-performance electron transport layers for perovskite solar cells on flexible pet versus on Glass’, ACS Applied Energy Materials, 5(4), pp. 4096–4107. doi:10.1021/acsaem.1c03311.

[134] Wang, L. et al. (2020) ‘Carrier Transport Composites with suppressed glass-transition for stable planar perovskite solar cells’, Journal of Materials Chemistry A, 8(28), pp. 14106–14113. doi:10.1039/d0ta03376f.

[135] Wu, W.-Q. et al. (2019) ‘Bilateral alkylamine for suppressing charge recombination and improving stability in blade-coated perovskite solar cells’, Science Advances, 5(3). doi:10.1126/sciadv.aav8925.

[136] Nakka, L. et al. (2022) ‘Analytical Review of spiro-ometad hole transport materials: Paths toward stable and efficient perovskite solar cells’, Advanced Energy and Sustainability Research, 3(8). doi:10.1002/aesr.202200045.

[137] Ernestas Kasparavicius, Marius Franckevičius, Vida Malinauskiene, Kristijonas Genevičius, Vytautas Getautis, and Tadas Malinauskas. Oxidized Spiro-OMeTAD: Investigation of Stability in Contact with Various Perovskite Compositions. ACS Applied Energy Materials 2021 4 (12), 13696-13705 DOI: 10.1021/acsaem.1c02375

[138] Guo, Y. (2022) ‘A novel organic dopant for Spiro-OMeTAD in high-efficiency and stable perovskite solar cells’, Frontiers in Chemistry, 10. doi:10.3389/fchem.2022.928712.

Abstract

With the rapid development of the so-called cloud storage. The practice of storing and sharing scientific experimental data on different internet platforms grew tremendously, which led to a significant accumulation of data. Until recently, this large quantity of data has been neglected due to the lack of effective techniques to gather knowledge and useful information from these data. Nevertheless, the progress achieved in the data-driven techniques in the past decade offered unique opportunities to extract important information from the material data. In this thesis, we have used advanced techniques of machine learning to solve critical problems that prevent successful commercialization of the perovskite solar cells technology through the investigation and the analysis of an important amount of data consisting of the measurements and materials information related to the manufacturing and the operation of these devices. This research provides a practical and useful guide for improving the performance of this kind of solar cell as well as enhancing the operational lifetime. Moreover, we have compared the results of machine learning with different previous experimental research, where we found remarkable coincidences. Which opens up important prospects in this field.

ملخص

مع التطور السريع لما يسمى بالتخزين السحابي. اصبح تخزين البيانات المستخلصة من التجارب العلمية ومشاركتها على منصات الإنترنت المختلفة شائعا جدا، مما أدى إلى تراكم كبير للبيانات. وحتى وقت قريب، تم إهمال هذه الكمية الكبيرة من البيانات بسبب عدم وجود تقنيات فعالة لجمع الخصائص والمعلومات المفيدة من هذه البيانات. ومع ذلك، فإن التقدم المحرز في التقنيات المعتمدة على البيانات في العقد الماضي أتاح فرصًا فريدة لاستخلاص معلومات مهمة من البيانات المتعلقة بفيزياء المواد. في هذه الأطروحة، استخدمنا تقنيات متقدمة من الذكاء الاصطناعي وبالتحديد للتعلم الآلي لحل العراقيل التي تمنع التسويق التجاري الناجح لتكنولوجيا الخلايا الشمسية من نوع البيروفسكايت من خلال استقراء وتحليل كمية مهمة من البيانات التي تحتوي على معلومات عن تصنيع وعمل هذه الأجهزة. ويقدم هذا البحث دليلاً عملياً ومفيداً لتحسين أداء هذا النوع من الخلايا الشمسية بالإضافة إلى تعزيز العمر التشغيلي لها. علاوة على ذلك، قمنا بمقارنة نتائج التعلم الآلي مع مختلف الأبحاث التجريبية السابقة، حيث وجدنا توافق ملحوظ بينهم. مما يفتح آفاقا مهمة في هذا المجال.

Résume:

Avec le rapide développement de ce qu'on appelle le « cloud stockage ». La pratique consistant à stocker et à partager des données scientifiques expérimentales sur des différentes plateformes d'internet est considérablement développée, ce qui a conduit à une accumulation importante de données scientifiques. Jusqu'à récemment, cette grande quantité de données a été négligée en raison du manque des techniques efficaces pour recueillir des informations à partir de ces données. Néanmoins, les progrès réalisés dans les "data-driven" techniques au cours de la dernière décennie ont offert des opportunités uniques pour extraire des informations importantes à partir des données des matériaux. Dans cette thèse, nous avons utilisé des techniques avancées de Machine Learning pour résoudre des problèmes critiques qui empêchent la commercialisation de la technologie des cellules solaires à pérovskite grâce à l'investigation et à l'analyse d'une quantité importante de données constituées avec des informations sur les matériaux liées à la fabrication de ces appareils. Cette recherche fournit un guide pratique et utile pour améliorer les performances de ce type de cellule solaire ainsi que sa durée de vie opérationnelle. De plus, nous avons comparé les résultats de Machine Learning avec plusieurs précédentes recherches expérimentales, dans lesquelles nous avons trouvé des coïncidences remarquables. Ce qui ouvre des perspectives importantes dans ce domaine.