People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research

IBN KHALDOUN UNIVERSITY OF TIARET

Dissertation

Presented to:

FACULTY OF MATHEMATICS AND COMPUTER SCIENCE
DEPARTEMENT OF COMPUTER SCIENCE

in order to obtain the degree of :

MASTER

Specialty: Software Engineering

Presented by:

ABBOU Mohamed salah eddine
GAFOUR Ahmed riadh

On the theme:

Modeling a photocatalytic reactor for water purification

Defended publicly on 12/06/2024 in Tiaret in front the jury composed of:

Mr BENDAOUD Mebarek	Pr	Tiaret University	Chairman
Mr AID Lahcene	MCA	Tiaret University	Supervisor
Mr MOSTEFAOUI Sid Ahmed	MCA	Tiaret University	Examiner

2023-2024

Acknowledgements

In the name of Allah, the Most Gracious, the Most Merciful. All praise is due to Allah, who has granted us the strength, patience, and knowledge to complete this thesis. We are deeply grateful for His guidance and blessings throughout this journey.

We extend our sincere thanks to our supervisor, Mr. Aid Lahcene, for his valuable guidance, constructive feedback, and continuous support throughout this research. His expertise and insights have been instrumental in shaping this work.

Our gratitude also goes to Mr. Dehbi Abdelkader and Mr. Bassaid Salah for generously providing the data that formed the core of this study. Their theoretical foundation on photocatalytic degradation has been a cornerstone of our research. Their contributions have greatly enhanced the quality and depth of this thesis.

May Allah reward all those who have supported us in this endeavor.

Abstract

Complex systems, characterized by their interconnected and dynamic components, can be effectively modeled using computational approaches, particularly neural network models. These models capture intricate patterns and behaviors within complex systems, which makes them powerful tools for modeling and prediction. Building on the strengths of neural networks, deep learning has demonstrated remarkable success across various domains due to its ability to learn different representations and model non-linear relationships. The aim of this work is to leverage deep learning techniques to model photocatalytic reactors, with the goal of optimizing the photocatalytic degradation process. We propose a model based on self attention mechanism for the prediction of the photocatalytic degradation rate and efficiency based on a set of experimental parameters. Base models are combined for data augmentation with a meta-model that incorporates a self-attention mechanism for prediction. The base models achieved excellent fits to the experimental data, and the meta-model attained a mean squared error of 0.0055 through five-fold cross-validation. tworks, deep learning has demonstrated remarkable success across various domains due to its ability to learn different represent

Keywords: photocatalysis, deep learning, self-attention mechanism, data augmentation, ensemble learning, persistent organic pollutants, small data.

ملخص

الأنظمة المعقدة، التي تتميز بمكوناتها المترابطة والديناميكية، يمكن نمذجتها بشكل فعال باستخدام الأساليب الحاسوبية، وبالأخص نماذج الشبكات العصبية. تلتقط هذه النماذج الأنماط والسلوكيات المعقدة داخل الأنظمة المعقدة، مما يجعلها أدوات قوية للنمذجة والتنبؤ. بالاستفادة من مزايا الشبكات العصبية، أظهر التعلم العميق نجاحًا كبيرًا عبر مختلف المجالات نظرًا لقدرته على تعلم تمثيلات متنوعة ونمذجة العلاقات غير الخطية. يهدف هذا العمل إلى الاستفادة من تقنيات التعلم العميق لنمذجة المفاعلات الضوئية، بهدف تحسين عملية التحلل الضوئي. نقترح نموذجًا يعتمد على آلية الانتباه الذاتي للتنبؤ بمعدل وكفاءة التحلل الضوئي بناءً على مجموعة من المعلمات التجريبية. تُدمج النماذج الأساسية من أجل زيادة البيانات مع نموذج ميتا يتضمن آلية الانتباه الذاتي للتنبؤ. حققت النماذج الأساسية توافقًا ممتازًا مع البيانات التجريبية، وحقق نموذج الميتا متوسط خطأ مربعات قدره 0.0055 من خلال التحقق المتقاطع بخمس طيات.

الكلمات المفتاحية: التحفيز الضوئي، التعلم العميق، آلية الانتباه الذاتي، زيادة البيانات، التعلم التجميعي، الملوثات العضوية الثابتة، البيانات الصغيرة

Contents

List of Figures

List of Tables

List of Abbreviations

POP	**P**ersistent **O**rganic **P**ollutant
TiO_2	**Ti**tanium Di**o**xide
MB	**M**ethelyne **B**lue
RSM	**R**esponse **S**urface **M**ethod
M&S	**M**odeling and **S**imulation
AI	**A**rtificial **I**ntelligence
ML	**M**achine **L**earning
DL	**D**eep **L**earning
ANN	**A**rtificial **N**eural **N**etwork
ReLU	**Re**ctified **L**inear **U**nit
tanh	**ta**nbolic **H**yperbolic
MSE	**M**ean **S**quared **E**rror
MAE	**M**ean **A**bsolute **E**rror
Adam	**Ad**aptive Moment Estim**a**tion

General introduction

0.1 Background

Unregulated industrialization and accelerated urbanization, driven by the rapid growth of the global population, have led to the discharge of a significant amount of toxic chemicals into water bodies [1]. Coupled with the escalating demand for clean water, this pollution has put a significant strain on many regions of the world. Estimates suggest that one in every four cities grapples with water stress [2].

Industrial waste, sewage leaks, and the widespread use of herbicides and pesticides in agricultural practices are the main sources of contamination for groundwater and surface water alike [3]. An annual extraction of 3,928 cubic kilometers of freshwater is reported, with agriculture and industry accounting for more than half [4, 5]. Apparel production alone contributes to 20% of industrial water pollution, with the dyeing process playing a prominent role [6, 7].

Persistent Organic Pollutants (POPs) are the underlying cause of many environmental and health concerns, mainly due to their disruptive nature and resistance to biodegradation [8]. POPs are lipophilic[1] and tend to accumulate in living organisms through processes of bioaccumulation and biomagnification [9]. This accumulation interferes with the physiological functions of the body and leads to hormonal abnormalities, reproductive and neurobehavioral disorders, and greater risks of cancer [10]. Moreover, POPs in water undergo a cyclical process of evaporation and deposition. They can travel large distances across bodies of water [11], contaminating the food chain and ultimately entering human consumption [12]. According to a study by the World Health Organization, improving access to clean water and sanitation services could avert up to 1.4 million deaths annually [13].

Various water treatment techniques have been developed to address this issue, with photocatalytic degradation standing as a promising solution [14]. It involves the use of photocatalysts, typically semiconductor materials like titanium dioxide (TiO_2), which, when exposed to ultraviolet (UV) or visible light, generate reactive oxygen species that can break down organic compounds such as pesticides [15] and dye-contaminated wastewater [16] into harmless byproducts [17]. Photocatalytic degradation has proven to be a sustainable and cost-effective technique in both experimental and practical use cases [18] and remains under active research.

However, experimentation in the field requires careful preparation and extensive control over interdependent parameters. Photocatalyst concentration, pollutant concentration, initial solution pH,

[1] Fat-soluble; tends to dissolve in fats, oils, and non-polar solvents.

reaction temperature, light intensity, and oxidant concentration are just a few of the experimental parameters that have a big impact on the photocatalytic degradation process [17]. Despite recent advances in computational photocatalysis and various approaches to simulate this process, the calculations employed often involve complex mathematical equations to describe the electronic structure and behavior of an already large molecular system. Some traditional simulations remain too complicated and computationally intensive to implement in practice [19].

With the current progress of machine learning (ML) approaches and the intricate and pattern-rich nature of chemistry, deep learning (DL) serves as a reliable method for modeling both the low-level processes and high-level outcomes of chemical reactions [20]. Numerous applications of DL have been suggested [21] in the fields of molecular design [22], reaction prediction [23], and drug discovery [24], to name a few.

Regarding photocatalysis, the growing research around attention mechanisms [25] is particularly compelling where the relations between experimental parameters dictate the efficiency of the degradation process [17]. Because experimental data is challenging to obtain and the range of potential conditions is vast, expert-guided data augmentation with the right experimental methodology can provide additional training samples while mitigating the effects of dataset bias.

In this work, we propose a deep learning model designed for small data, formulated within an ensemble learning framework and incorporating a self-attention mechanism during the integration phase, for the prediction of the photocatalytic degradation rate and efficiency from a set of experimental parameters.

0.2 Problem statement

The photocatalytic degradation of persistent organic pollutants in wastewater shows promise as a sustainable treatment method, but optimizing the process remains challenging due to the complex interplay of numerous experimental parameters. Extensive experimental trials are required to evaluate the effects of factors like photocatalyst loading, light intensity, and initial pollutant concentration on the degradation rate and efficiency. However, such experiments are laborious and time-consuming.

This research aims to develop a data-driven approach using deep learning to accurately model and predict the photocatalytic degradation performance from key experimental parameters, while minimizing the need for additional experimental data through novel data augmentation and ensemble techniques. The proposed model can guide the design of optimized photocatalytic systems and inform future experimental efforts in this critical area of environmental remediation.

0.3 Delimitation

The scope of this research was confined within certain boundaries to maintain a focused investigation. The key delimitations are as follows:

- **Photocatalyst material:** the study focused specifically on the photocatalytic degradation using a TiO_2/curcumin nanocomposite material deposited on a cellulose paper substrate. Other photocatalyst materials were not investigated.
- **Target pollutant:** methylene blue dye was used as the model persistent organic pollutant for the degradation experiments. The applicability of the models to other pollutant types was not explored.

Additionally, choices were made to narrow the study to specific experimental conditions and variables:

- **Experimental parameters:** only three key experimental parameters - photocatalyst mass, light intensity, and initial pollutant concentration - were varied and modeled. Other potential factors like solution pH, temperature, oxidant concentration, etc., were not considered.
- **Experimental design:** the data was obtained from experiments employing a one-factor-at-a-time variation approach. Interaction effects between multiple simultaneously varying parameters were not captured.

The modeling approaches were also delimited in terms of the techniques and data employed:

- **Model architecture:** the deep learning models were limited to specific neural network architectures, primarily using dense layers and a self-attention mechanism. Other architectural choices were not comprehensively evaluated.
- **Data availability:** the study operated under a limited initial experimental dataset, necessitating data augmentation techniques. The performance with larger datasets was not assessed.

0.4 Approach

The approach involved using Bouazza et al.'s [26] experimental data on the photocatalytic degradation of methylene blue dye using a TiO_2/curcumin nanocomposite catalyst under varying conditions of photocatalyst mass, light intensity, and initial pollutant concentration. The degradation kinetic data was preprocessed by fitting an exponential decay function to the data to extract rate parameters as modeling targets. To overcome the limited dataset size, the data was partitioned based on varying one experimental parameter at a time. A set of base deep neural network models were trained on these partitioned subsets to model localized degradation behaviors. These base models then generated augmented synthetic data over an expanded parameter range through predictive degradation curves. A meta deep neural network architecture incorporating a self-attention mechanism integrated information from the different experimental parameters and augmented data. This meta model was trained on the augmented dataset to learn a generalized mapping between parameters and predicted degradation profiles. After evaluating performance, the meta model identified optimal experimental parameter ranges maximizing photocatalytic degradation rate and efficiency.

This thesis is structured into three chapters:

- **Chapter 1: modeling and simulation.** This chapter serves as the foundation for understanding the principles and practices of modeling complex systems. It introduces the concept of modeling and simulation (M&S) and discusses the classification of systems into linear and nonlinear categories, as well as Zeigler's classification based on time characteristics. The chapter also explores various aspects of modeling, including computational models and the concepts behind the theory of M&S, such as levels of system specification and framework of M&S.
- **Chapter 2: fundamentals and techniques in machine learning and deep learning.** In this chapter, we introduce the basic principles of ML, covering artificial intelligence (AI), ML categories, and tasks, along with their applications in chemistry. Furthermore, we discuss the foundational concepts of DL, including artificial neural networks (ANN), key concepts, and attention mechanisms, and provide a comprehensive understanding of the underlying principles that support these techniques.
- **Chapter 3: Photocatalytic degradation model.** This chapter introduces a novel approach to enhance the performance of DL models in predicting photocatalytic degradation. It outlines the methodology used, which involves implementing a data augmentation strategy. The chapter details the process of partitioning the data, training base deep neural network models, and constructing a meta model with a self-attention mechanism. It concludes with a discussion of the results obtained and the implications for optimizing experimental parameters to maximize degradation efficiency.

Chapter 1

Modeling and simulation

1.1 Introduction

M&S techniques play a vital role in understanding and analyzing complex systems across numerous disciplines. From engineering to economics, biology to physics, these computational tools allow researchers to create virtual representations of real-world phenomena, which facilitates their exploration, prediction, and optimization. This chapter covers the classification of systems, computational models, and the underlying theory of M&S. It discusses approaches like the classification of systems into linear and nonlinear models, as well as continuous and discrete models. The theory of M&S covers topics such as model verification, validation, and analysis techniques.

1.2 Complex systems

A complex system can be described as an intricate network comprised of numerous components or elements that interact with one another. These interactions often occur in a nonlinear manner, they deviate from simple linear equations or dynamics. Complex systems are neither entirely orderly nor completely disordered, but rather exhibit emergent behaviors and patterns that manifest at larger scales through a process of self-organization. In other words, the collective interactions among the system's components give rise to order and structure that cannot be easily predicted or explained by studying the components in isolation [27].

Another way to think about a complex system is as a set of crucial attributes and necessary conditions. As previously stated, a complex system is made up of many parts or components that interact with one another in intricate and interconnected ways. These complex relationships result in feedback loops, where over time, the actions of one component both influence and are influenced by the actions of other components. The system displays spontaneous order and patterns that result from the collective interactions among the components, even in the absence of centralized control. This emergent order is hierarchically structured, with multiple levels of organization and properties that interact across levels, both higher and lower. Nonlinearity, although common, is not a prerequisite for complexity. Additionally, emergence alone does not fully define complexity. Instead, a complex system is characterized by factors such as numerosity, intricate interactions, feedback loops, spontaneous order, hierarchical organization, and robustness [28].

1.3 Classification of systems

Systems exhibit diverse characteristics, and researchers have classified them based on various factors.

1.3.1 Behavior-based classification

1.3.1.1 Linear systems

Linear systems are characterized by a direct proportionality between their inputs and outputs, this means that any variation in the input induces a proportionate alteration in the output [29]. Such systems can be modeled using linear equations or transfer functions. Examples include basic electrical circuits, and specific chemical reactions. The primary benefit of linear systems lies in their relative simplicity in terms of analysis and control, attributable to their predictable behavior [30].

1.3.1.2 Nonlinear systems

Nonlinear systems do not adhere to a linear relationship between inputs and outputs [29]. Their behavior can be intricate and may involve nonlinear equations or differential equations. Nonlinear systems can exhibit various phenomena, such as multiple equilibrium points, limit cycles, and chaotic behavior. Analyzing and controlling nonlinear systems is more challenging due to their complexity [30]. Examples of nonlinear systems include weather patterns, biological ecosystems.

1.3.2 Zeigler's classification

According to Zeigler [31], complex systems can be classified into Continuous-Time Systems, Discrete-Time Systems, Discrete Event Systems.

1.3.2.1 Continuous-time systems

Continuous-time systems operate without interruption over time, characterized by continuous functions. These systems find applications in diverse fields, including physics, engineering, and natural phenomena. For instance, analog electrical circuits obey continuous-time dynamics, described by differential equations. Fluid dynamics, heat transfer, and chemical reactions also fall within this category.These processes can be modeled using partial differential equations.

1.3.2.2 Discrete-time systems

Discrete-time systems process signals at specific time intervals. They play a crucial role in digital signal processing, control systems, and communication networks. Key features include sampling instants and discrete values. Digital filters, such as finite impulse response (FIR) and infinite impulse

response (IIR) filters, rely on discrete-time modeling. Zeigler's contributions cover the discrete event system specification (DEVS) formalism, which encompasses discrete-time modeling.

1.3.2.3 Discrete event systems

Discrete event systems exhibit behavior at arbitrarily spaced moments, triggered by relevant events. These events cause state transitions. DEVS provides a rigorous foundation for modeling and simulating such systems. Examples include queuing networks, communication protocols, and manufacturing processes. Zeigler's work [31] on DEVS addresses concurrency, synchronization, and event-driven interactions, which makes it indispensable in this domain.

1.4 Modeling

1.4.1 Introduction to modeling

Modeling is a fundamental tool in various scientific and engineering fields, it serves as a language of the universe and a tool for the study of patterns of all kinds. It makes the simulation and testing of complex systems possible, and provides insights that might not be attainable through experimentation alone, and has proven to be fundamental to design and optimize systems [32]. For example, in environmental science, mathematical models can predict the consequences of climate change.

1.4.2 Principles and practices of modeling

Modeling is a fundamental tool in scientific and engineering disciplines [33] to encapsulate our understanding of a system and predict its behavior under different conditions. At their core, models are approximations of reality, designed to provide an approximate description that is good enough for the intended purpose. The level of approximation can vary depending on the specific requirements of the model, but models are inherently simplified representations of complex real-world systems.

The process of constructing models involves adherence to foundational principles such as Occam's Razor (simplicity), conservation laws (physical consistency), causality and mechanistic understanding (capturing causal relationships), predictive accuracy, flexibility, and adaptability. The principle of abstraction is a key aspect of this process, where complex real-world systems are distilled into simplified representations that capture essential features while they disregard irrelevant details, which helps to focus on the core dynamics of the system under study.

Validation and verification are essential steps in assessing the accuracy and reliability of models [34]. Validation involves comparing model outputs with empirical data to assess their predictive accuracy, while verification ensures that the model accurately represents the underlying theory or conceptual framework. A model validated under one set of conditions may not necessarily be valid under different conditions. This highlights the importance of continuous evaluation and refinement.

In scientific research and engineering, the process of modeling is iterative [35]. Initially, models are constructed with simplicity; they often rely on limited data and a basic understanding of the phenomenon under study. However, as additional data becomes available and our comprehension deepens, these models undergo refinement and updates. This iterative cycle of model development and enhancement constitutes a crucial component of the scientific method. This implies that models predictive capabilities are improved through iterative refining.

The ultimate test of a model is its predictive power—the ability to make accurate predictions about the system's behavior under new conditions or scenarios. Additionally, models serve as a means of communication; they provide a formalized representation of our understanding that can be shared, critiqued, and built upon by others in the field.

1.4.3 Computational models

Computational modeling, a multidisciplinary approach that integrates mathematics, physics, and computer science, employs computer systems to simulate and scrutinize intricate systems. A computational model is comprised of several variables that encapsulate the attributes of the system. The process of simulation involves the manipulation of these variables, either individually or collectively, to observe the resultant effects. This computational approach allows the execution of a multitude of simulated experiments digitally. The vast array of digital experiments serves to pinpoint a select few laboratory experiments that hold the highest potential for resolving the research question at hand [36].

1.5 Theory of modeling and simulation

The theory of M&S establishes a structured approach to develop models that can represent real-world systems and simulate their behavior. This theory defines different levels of system specification, ranging from low-level observations to high-level compositional structures. It uses methodologies such as Klir's knowledge levels and the hierarchy of system specifications to methodically construct models across multiple abstraction layers [31].

The M&S theory delineates key concepts such as the source system, experimental frames, model components, simulators, and techniques for model validation. It allows the development of robust models that can replicate, predict, and outline the detailed dynamics of complex systems.

1.5.1 Levels of system specification

1.5.1.1 Klir's knowledge levels

Klir's Knowledge Levels propose a structured framework to comprehend the various layers of information and understanding that can be acquired about a system. This hierarchy defines four distinct levels of knowledge [37].

The source level Represents the most basic level of system knowledge. The goal at that foundational level is to pinpoint the precise real-world phenomenon or entity that will be modeled, together with the methods that will be used for observation and measurement. At the source level, it is necessary to identify the relevant variables that need to be observed and the methods that will be used to measure and record their values over time.

The data level The collection of empirical measurements and observations obtained from the source system becomes the main emphasis at the data level. This stage includes gathering unprocessed data, which serves as the foundation of evidence upon which subsequent levels of understanding are constructed.

The generative level The generative level represents a change from the simple collection of data to being able to reproduce and regenerate the phenomena that have been observed by using small mathematical or computational representations. In order to capture the generative mechanisms that underlie the behavior of the system, models or algorithms that can replicate the patterns in the observed data are developed at this level.

the structure level The structure level represents the most comprehensive and intricate level of system understanding. The objective is to clarify the internal compositional structure of the system as well as the interrelationships and interactions among its constituent components. The complex network of interdependent subsystems and their collective dynamics, which collectively explain the system's observed behavior, are revealed at this level.

1.5.1.2 Klir's systems problems

Klir's framework highlights three fundamental problems that arise when analyzing and manipulating systems. These problems are intricately tied to the transitions between different levels of system knowledge and underscore the complexity inherent in the study and manipulation of systems [37].

Systems analysis This problem concerns the understanding of a system's behavior based on a priori knowledge of its structural composition. It entails the transfer of information from higher levels of system knowledge, specifically the structure level, to lower levels, such as the data level or the generative level. The main goal of systems analysis is to deduce the observable manifestations and output patterns that arise from the intricate interplay of the system's constituent components and their interconnections.

Systems inference Systems inference, as opposed to systems analysis, addresses the inverse challenge, which is the deduction of the system's underlying structure given only the observation of its behavior. This task requires the transition from lower levels of system knowledge, usually the data level, to higher levels, notably the generative level or the structure level. The primary objective is

to reconstruct the internal compositional architecture and governing mechanisms that produce the observed data patterns, which allows for a deeper understanding and the potential to manipulate the system effectively.

Systems design The problem of systems design combines difficulties found in both systems analysis and systems inference. It revolves around the conceptualization and construction of novel systems or the redesign of existing ones, all driven by specific desired behaviors or output patterns. Systems design necessitates the integration of knowledge across multiple levels, beginning with lower levels (e.g., data level or generative level) and culminating in the formulation of a structure level specification that can realize the intended system functionality.

1.5.1.3 Hierarchy of system specifications

The Hierarchy of System Specifications, offers a structured and systematic approach to construct models and establish relationships across different levels of system specifications. The five different specification levels defined by this hierarchical paradigm gradually capture the subtleties of a system's behavior and structure. This unified framework facilitates the representation and analysis of systems across diverse domains and abstraction levels and enables the integration of various modeling paradigms to develop robust M&S methodologies [38].

Level	Specification name	Corresponds to Klir's	What we know at this level
0	Observation frame	Source system	How to stimulate the system with inputs; what variables to measure and how to observe them over a time base.
1	I/O behavior	Data system	Time-indexed data collected from a source system; consists of input/output pairs.
2	I/O function		Knowledge of initial state; given an initial state, every input stimulus produces a unique output.
3	State transition	Generative system	How states are affected by inputs; given a state and an input, what is the state after the input stimulus is over; what output event is generated by a state.
4	Coupled component	Structure system	Components and how they are coupled together. The components can be specified at lower levels or can even be structured systems themselves – leading to hierarchical structure.

TABLE 1.1: Relation between system specification hierarchy and Klir's levels [31]

Observation frame At this foundational level, the specification focuses on defining the means by which the system will be stimulated and observed. This level entails the identification of the system's input and output interfaces, the variables to be measured, and the temporal domain over which observations will be made. The Observation Frame establishes the experimental context and the boundaries within which the system's behavior will be explored.

Input/Output (I/O) behavior This level encapsulates the collection of input/output pairs gathered through observation. It represents the empirical data that captures the system's external manifestations, without delving into the underlying mechanisms that govern its behavior.

I/O function The I/O Function introduces the concept of initial state knowledge, which enables the establishment of a functional relationship between the input and output trajectories. At this level, the specification incorporates information about the system's initial conditions, allowing for the prediction of unique output responses based on the provided inputs and the known initial state.

State transition At this stage the specification expands to encompass not only the initial state but also the dynamics that govern the system's evolution over time. The state transition mechanism is defined, which dictates how the system's internal state is transformed in response to external inputs, as well as the state-to-output mapping, this leads to the determination of the observable outputs generated by the system based on its current state.

Coupled component The Coupled Component specification provides a comprehensive representation of the system's internal structure. This level delineates the constituent components that comprise the system and the intricate couplings and interconnections among them. The components themselves can be specified at lower levels of the hierarchy, to construct hierarchical and modular system representations.

1.5.2 Framework of M&S

1.5.2.1 Source system

The source system, whether real or artificial, is the environment or process of interest in the modeling effort. It is regarded as a source of observable data, which is collected as time-indexed trajectories of variables. These trajectories are compiled into a system behavior database, which represents the empirical data obtained from observing or experimenting with the system. This database is essential for the development and validation of models, as it provides the empirical foundation for simulation efforts. [31]

The data within the system behavior database is acquired through experimental frames, which define the specific conditions and variables of interest to the modeler. These frames ensure that the data collection process is systematic and aligned with the study's objectives. The quality and quantity of

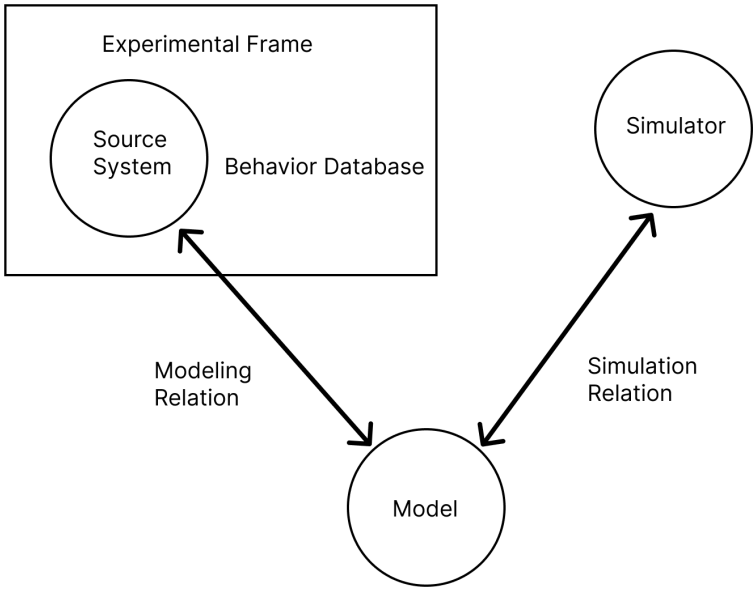


FIGURE 1.1: Modeling and simulation framework [31]

data available to populate the system database can vary significantly across different applications of M&S. In data-rich environments, abundant prior experimentation or readily obtainable measurements provide ample data. Conversely, data-poor environments may offer limited historical data or data of questionable representativeness. In such cases, it may be challenging or expensive to acquire better data, necessitating a targeted approach to data collection driven by the modeling process.

In certain situations, the acquisition of high-quality data might be impossible, due to various constraints, while in other contexts, the process can be prohibitively expensive. This means that model accuracy and reliability are enhanced when data collection is strategically focused on important areas.

1.5.2.2 Experimental frame

The experimental frame specifies the conditions that dictate the interaction between the system of interest and the observer. It encompasses the variables and parameters essential to accurately capture the behavior of the system under specific scenarios. The reason this operationalization is important is that it allows modelers to focus on particular aspects of the system that are relevant to their objectives. For instance, in modeling a forest fire, an experimental frame might include variables such as lightning strikes, rain, wind, and smoke. More refined frames could incorporate additional factors like the moisture content of vegetation and the quantity of unburned material [39].

Components of an Experimental Frame An experimental frame consists of three primary components that work in unison to facilitate the observation and experimentation of the system, and ensure that the data collected is relevant and precise.

- **Generator:** this component is responsible for the production of input segments for the system. It defines the stimuli or conditions under which the system will be tested or observed.
- **Acceptor:** the acceptor monitors the experiment to ensure that the desired conditions are met. Serving as a gatekeeper, it verifies that the experimental setup aligns with the specified frame.
- **Transducer:** this component observes and analyzes the system's output segments. It converts the raw data generated by the system into a form that can be interpreted and utilized to achieve the project's objectives.

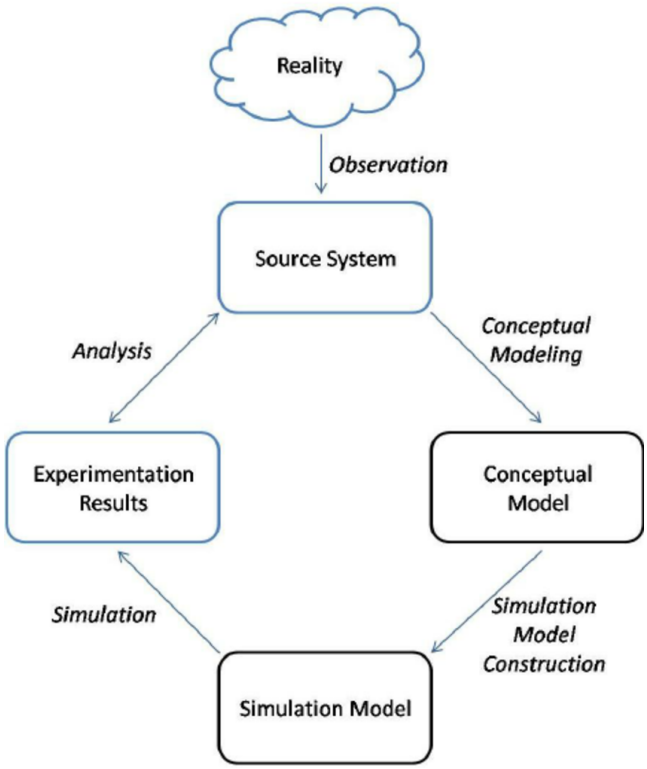


FIGURE 1.2: Modeling and simulation lifecycle [40]

1.5.2.3 Model

In the M&S framework, a model is a crucial entity that represents a system specification designed to generate Input/Output (I/O) behavior. This specification often takes place at the structural and generative levels [31].

A model can be understood as a set of rules, instructions, equations, or constraints that guide the generation of behavior from a system. The model can process input trajectories and generate corresponding output trajectories based on its initial state with the use of methods for state transitions and output generation. The model's main goal is to encapsulate the dynamic aspects of the system it represents.

State transition mechanisms These define how the model changes state in response to inputs over time. State transitions are critical to capture the system's dynamic behavior and ensure that the model can accurately replicate the sequence of changes that accrues within the actual system.

Output generation mechanisms These mechanisms specify how the model generates outputs from its current state and inputs. This is vital on order to produce observable behaviors that can be compared with the real system's behavior during validation processes.

Validation of models The validity of a model is determined by its ability to faithfully represent the real system within a specified experimental frame. This involves several types of validity:

- **Replicative validity:** ensures that the model can reproduce the observed behavior of the system under the same experimental conditions. This is the most basic form of validity and is confirmed when the model's outputs agree with the system's outputs within acceptable tolerances across all possible experiments within the frame [41].
- **Predictive validity:** requires the model not only to replicate known behaviors but also to accurately predict new, unseen behaviors of the system. This type of validity is necessary for models used in forecasting and future scenario analysis and involves agreement at the I/O function level of the system hierarchy.
- **Structural validity:** this is the highest level of validity, ensuring that the model not only mimics the system's output but also replicates the internal processes and transitions of the system. This involves agreement at the state transition or coupled component level, demonstrating that the model accurately represents the system's structure and internal dynamics.

Accuracy and fidelity Accuracy refers to the degree of closeness of the model's outputs to the real system's outputs. Fidelity, on the other hand, encompasses both the detail and the validity of the model. A high-fidelity model is expected to be detailed and valid, able to accurately capture the nuances of the system's behavior. However, it is crucial to note that high detail alone does not guarantee high validity; a model can be detailed yet inaccurate if it misrepresents the system's functioning at a fundamental level.

1.5.2.4 Simulator

A model, represented as a set of instructions, necessitates an agent that is able to execute those instructions and produce the corresponding behavior. This agent is referred to as a simulator.

A simulator is defined as a computational system, which could be a single processor, a network of processors, the human mind, or even an abstract algorithm. Its primary function is to execute a given model, so it generates the expected behavior. Generally, simulators are specified at a high level and designed using well-understood, off-the-shelf components.

1.5.2.5 Model and simulator separation

The separation of the model and simulator concepts offers several advantages within the framework:

- **Model portability and interoperability:** it becomes possible to run the same model with several simulators by expressing it in a certain formalism. At a higher abstraction level, this facilitates portability and interoperability.
- **Simulator correctness:** we can confirm the validity of the algorithms developed for simulators aligned with various formalisms and make sure the simulator accurately simulates the behavior of the model.
- **Model complexity measurement:** the resources required by a simulator to faithfully execute a model serve as a valuable measure of the model's complexity.

1.6 Conclusion

In conclusion, M&S provides a structured framework to represent and analyze complex systems across various domains. The M&S theory establishes a hierarchy of system specifications; it ranges from observational frames to coupled component structures. The framework delineates the interplay between source systems, experimental frames, models, and simulators for the development of valid models that accurately capture system behavior. The separation of models and simulators promotes model portability, simulator correctness, and complexity measurement. M&S offers a powerful methodology that helps understand, predict, and manipulate complex systems.

Chapter 2

Fundamentals and techniques in machine learning and deep learning

2.1 Introduction

In recent years, the emergence of advanced Machine Learning techniques has opened up new possibilities for accelerating progress in various subfields of chemistry. Machine learning, which encompasses a range of algorithms and statistical models that enable computers to improve their performance on tasks through experience, has shown remarkable potential in handling the complexity and vast data sets. A notable development within this domain was the evolution of artificial neural networks into more sophisticated DL architectures, which have since revolutionized numerous aspects of the field, from molecular design to reaction prediction.

2.2 Foundations of Deep Learning

Deep learning is an advanced subset of machine learning, where algorithms autonomously learn from data patterns without explicit instructions. Central to this approach are artificial neural networks, which are inspired by the structure and functionality of the human brain.

2.2.1 Core Concepts of Machine Learning

Machine learning is founded on the idea that systems can evolve and enhance their performance by processing data, without being directly programmed. This involves using algorithms that can analyze, model, and derive insights from data, facilitating decision-making and predictive capabilities.

2.2.1.1 Artificial Intelligence

Artificial intelligence (AI) refers to the creation of systems that emulate human cognitive functions, allowing machines to perform tasks that traditionally required human intelligence. AI systems can learn from experience, identify patterns, and adapt to new scenarios [42, 43]. This adaptability makes

AI indispensable in various fields, including recommendation systems [44], civil engineering [45], healthcare [46], and finance [47].

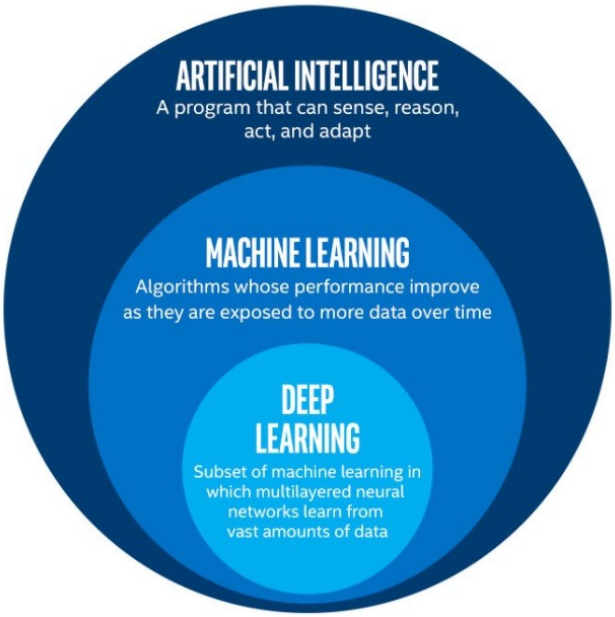


FIGURE 2.1: Exploration of machine learning techniques

AI is omnipresent in our modern society. Wherever expert systems are in place, it is highly likely that an AI algorithm is being used to improve performance and accuracy.

2.2.1.2 Applications in chemistry

AI techniques have been successfully applied to a wide range of problems and domains of chemistry. In bioinformatics and computational biology, techniques like Deep Belief Networks (DBNs) and Deep Boltzmann Machines (DBMs) have been employed for tasks such as protein structure prediction and drug discovery [48].

The attention mechanism has been applied across various domains in chemistry which led to significant advancements in the field. For instance, attention mechanisms have been employed in chemical reaction prediction. The study by Su et al. [49] highlights the use of self-attention in predicting the feasibility of copper(I)-catalyzed alkyne–azide cycloaddition reactions. The model not only enhances prediction accuracy but also provides insights into underlying reaction mechanisms [49].

2.2.1.3 Machine learning

Machine Learning is a subdiscipline of AI (see figure 2.1) that enables machines to learn from data without being explicitly programmed. In other words, instead of writing a specific program to perform a given task, a machine learning algorithm can be trained on a dataset to identify patterns

and relationships in this data, then use this knowledge to make predictions or decisions on new data not previously explored by the algorithm [43].

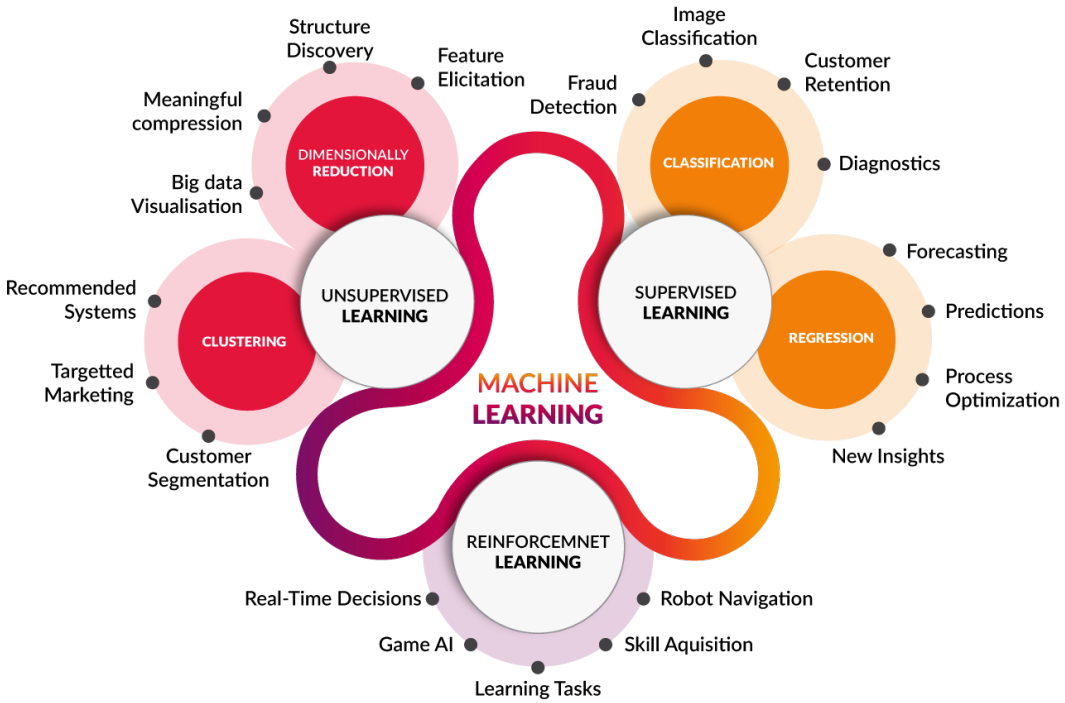


FIGURE 2.2: Some applications and types of machine learning / Christelle Julias, smartpredict.ai

Definition 1 (Machine Learning)**.** *Machine learning involves creating algorithms that improve their performance on a specific task by learning from data or experiences over time.*

The *experience* that a machine learning algorithm learns from can take various forms: labeled or unlabeled data, user feedback, logs or census archives, or interactions with the environment.

2.2.1.4 Categories of machine learning

Machine learning methods fall into four main categories, including semi-supervised learning, depending on the task and the type of feedback available to the learning system [43]:

- **Supervised learning:** The algorithm receives a training dataset with input/output pairs, where input features are represented as vectors, and output labels are predefined categories or values [43, 50]. The goal is to learn a function that accurately predicts the output label for new inputs. Examples include image classification, speech recognition, natural language processing, and recommendation systems.
- **Unsupervised learning:** In this category, the algorithm identifies patterns in the data without labeled examples. It receives an input dataset without corresponding output labels, aiming

to discover meaningful patterns or structures like clusters, principal components, or frequent patterns [50]. Applications include anomaly detection, data compression, and data visualization.

- **Semi-supervised learning:** The algorithm works with both labeled and unlabeled examples, assuming that labeled examples are more costly or harder to obtain. The goal is to use the labeled examples to guide the learning process and enhance the model's accuracy on unlabeled data [50].
- **Reinforcement learning:** The agent[1] learns to take actions in an environment to maximize a cumulative reward signal. Feedback comes as rewards or penalties based on actions, with the objective of learning a policy that maximizes the expected cumulative reward over time [43, 50].

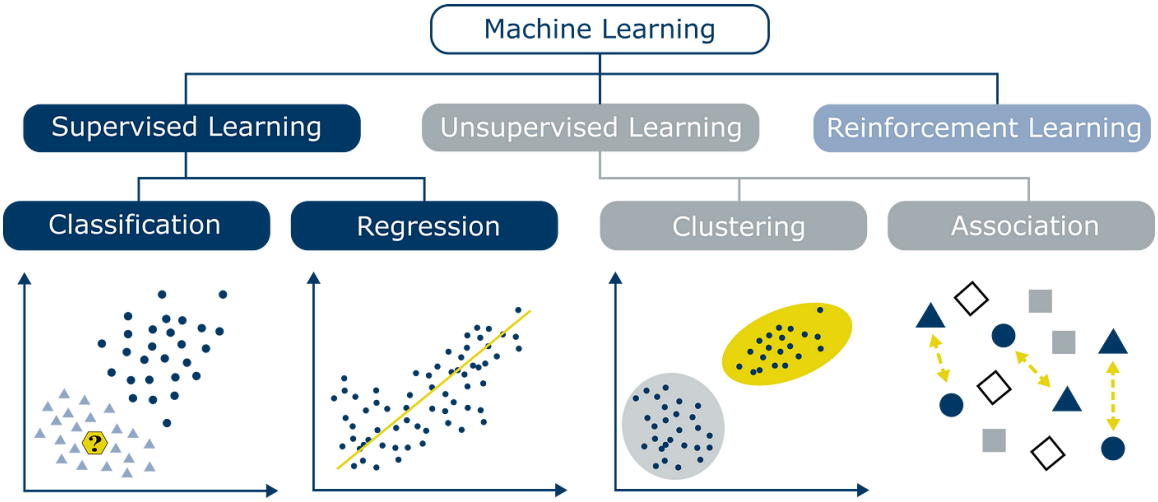


FIGURE 2.3: The most common machine learning tasks / Dominik Polzer, Towards-DataScience

2.2.1.5 Machine learning tasks

Machine learning algorithms apply to a variety of tasks, with common examples being **regression**, **classification**, and **clustering** (see Figure 2.3) [43, 50].

- **Regression:** The computer receives input data and predicts a numerical value, i.e., a function f that maps input data to a numerical output[2] $f : \mathbb{R}^n \to \mathbb{R}$. For example, predicting a house's price based on its characteristics such as size, number of bedrooms, or location [43].
- **Classification:** The goal is to predict the class or category of a new data point based on its features. The algorithm trains on labeled examples, and the output is a function mapping

[1] An agent perceives its environment through sensors and acts on it through actuators to achieve specific goals.
[2] Here, n represents the dimensionality of the feature vector.

input data to a discrete output[3] [43]. For instance, identifying whether an email is spam or not, which involves two classes: *spam* and *not spam*, making it a binary classification.

- **Clustering:** This involves grouping similar data points based on their features. The objective is to partition data into distinct clusters where points in each group are more similar to each other than to those in other groups. The algorithm discovers the underlying structure of the data without labeled examples [50].

2.2.1.6 Limitations of machine learning

Despite significant progress, several challenges remain to make machine learning more effective and reliable:

- **Data quality and availability:** Data may be incomplete, contain outliers, or be biased, affecting prediction quality. Careful cleaning, preprocessing, and strategies to handle missing data are essential [51].
- **Interpretability and explainability:** Interpretability is understanding how a model makes predictions, while explainability provides clear reasons for these predictions [52]. Complex models can be opaque, raising concerns in critical areas like healthcare.
- **Bias and fairness:** Bias arises when training data contains systematic errors or is unrepresentative, leading to unfair predictions for certain groups [53].

Addressing these challenges is crucial for the advancement and broader adoption of machine learning across various domains and industries.

2.2.2 Introduction to artificial neural networks

Artificial neural networks (ANNs) are computational models inspired by the human brain. They address complex problems where relationships between features are not easily explained.

2.2.2.1 Functioning of an artificial neural network

A neural network is composed of several layers of interconnected neurons. Each neuron receives weighted inputs, combines them using an activation function, and produces an output [43, 54]. It is common to illustrate a neural network in two ways: the sum of products method or the linear transformation method (see figure 2.4). In the first approach, linear operations are broken down into a series of simpler products and sums. In the second approach, linear transformations are considered directly. These two methods allow representing the operations performed by a neural network, but they are equivalent in terms of functionality. Formally, let X be the set of inputs (input vector), W the

[3] A discrete value, or categorical, is a label representing a class.

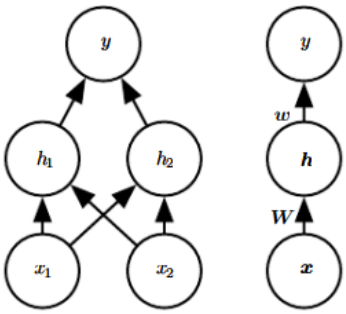


FIGURE 2.4: Among the ways to illustrate a neural network [55]

weight matrix that combines the weighted inputs, b the bias vector, and σ the activation function. The output Y of a neuron layer is given by

$$Y = \sigma(WX + b), \tag{2.1}$$

$WX + b$ represents the linear transformation of the inputs, followed by the application of the activation function σ. The parameters W and b are optimized through optimization algorithms such as gradient descent (see paragraph 2.2.2.3) during learning. For example, for the matrix dimensions illustrated in equation (2.2), the output Y will be of dimension $\mathbb{R}^2$,

$$X = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \quad W = \begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}. \tag{2.2}$$

Interestingly, if the number of rows in matrix W and vector b were equal to 1, the neural network model would reduce to a linear regression (see figure 2.5). In this case, the output Y would simply be a linear combination of the inputs weighted by a single weight and added with a single bias:

$$Y = \sigma(w_{11}x_1 + w_{12}x_2 + w_{13}x_3 + b_1). \tag{2.3}$$

The activation function σ would then be applied to this linear combination. However, the power of artificial neural networks lies in their ability to learn non-linear relationships and expand dimensionality to accommodate more information and offer a certain flexibility [55].

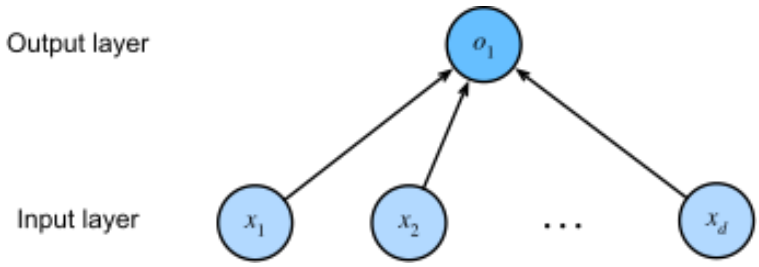


FIGURE 2.5: Linear regression represented as a neural network

2.2.2.2 Context and history of artificial neural networks

Because linear regression predates computational neuroscience, it may seem anachronistic to describe linear regression in terms of neural networks. Nevertheless, they were a natural starting point when cyberneticists and neurophysiologists Warren McCulloch and Walter Pitts began developing artificial neuron models [56].

Consider the illustration of a biological neuron in figure 2.6, composed of dendrites (input terminals), the nucleus (CPU), the axon (output wire), and axon terminals (output terminals), allowing connections with other neurons via synapses [57].

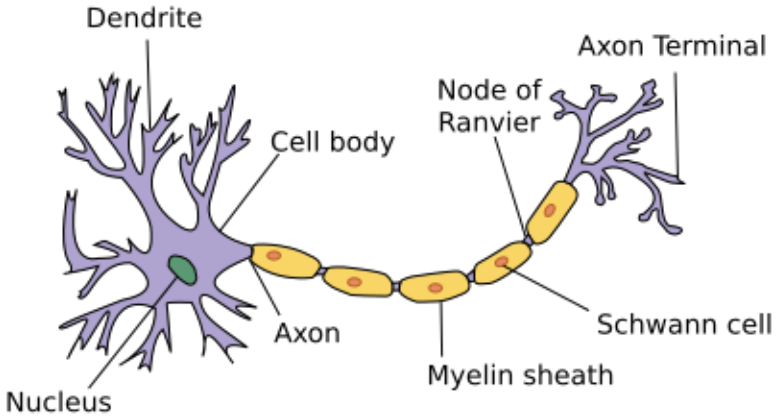


FIGURE 2.6: Illustration of a biological neuron [57]

Information x_i from other neurons (or environmental sensors) is received in the dendrites. Specifically, this information is weighted by synaptic weights w_i, determining the effect of the inputs, for example, activation or inhibition via the product $x_i w_i$. The weighted inputs from multiple sources are aggregated in the nucleus as a weighted sum $y = \sum_i x_i w_i + b$, possibly subjected to non-linear post-processing via a function $\sigma(y)$. This information is then sent via the axon to the axon terminals, where it reaches its destination[4] or is transmitted to another neuron via its dendrites [56].

Certainly, the high-level idea that many units of this type could be combined, given the right connectivity and learning algorithm, to produce behavior far more interesting and complex than what a single neuron could express, stems from studies of real biological neural systems.

However, most of today's deep learning research draws inspiration from a much broader source. We refer to Russell and Norvig [54] who pointed out that, while airplanes may have been inspired by birds, ornithology has not been the main driver of aeronautical innovation for a few centuries. Similarly, inspiration in deep learning comes equally, if not more so, from mathematics, linguistics, psychology, statistics, computer science, and many other fields.

[4]For example, an actuator such as a muscle.

2.2.2.3 Learning and optimization of a neural network

Naturally, fitting a neural network to data requires that we agree on a measure of fitness (or, equivalently, unfitness). Loss functions quantify the distance between the actual and predicted values of the target [43, 54].

Loss function The loss[5] will generally be a non-negative number where smaller values are better and perfect predictions result in a loss of 0. For regression problems, the most common loss function is the squared error (equation 2.4). When our prediction for an example i is $\hat{y}_i$ and the corresponding actual label is y_i, the squared error is given by

$$l_i(W, b) = \frac{1}{2}(\hat{y}_i - y_i)^2, \tag{2.4}$$

the constant $\frac{1}{2}$ makes no real difference, but it proves convenient for notation, as it cancels out when we take the derivative of the loss (which will be relevant in the next paragraph 2.2.2.3). Since the training dataset is given to us and is therefore out of our control, the empirical error depends only on the model parameters W and b.

To measure the quality of a model on the dataset of n examples, we simply take the average[6] (or, equivalently, the sum) of the losses over the training set [43, 50, 55]:

$$L(W, b) = \frac{1}{n}\sum_{i=1}^{n} l_i(W, b) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2}\left(WX_i + b - y_i\right)^2. \tag{2.5}$$

Optimization with the gradient descent algorithm While simple problems like linear regression may admit analytical solutions, this is unfortunately not always the case. Analytical solutions allow for nice mathematical analysis, but the requirement for an analytical solution is so restrictive that it would exclude almost all aspects of deep learning [50, 55].

Fortunately, even in cases where we cannot solve models analytically, we can often, in practice, still train models efficiently. Moreover, for many tasks, these difficult-to-optimize models prove to be so much better that figuring out how to train them is well worth it [55].

The key technique for optimizing almost all deep learning models is to iteratively reduce the error by updating the parameters in the direction that progressively decreases the loss function. This algorithm is called gradient descent.

In its most basic form[7], at each iteration t, we first randomly sample a minibatch $\mathcal{B}_t$ composed of a fixed number $|\mathcal{B}|$ of training examples. Then, we calculate the derivative (gradient) of the average loss on the minibatch with respect to the model parameters. Finally, we multiply the gradient by a

[5] The term *cost* can also be found referring to it.

[6] In this case, it is the mean squared error defined in equation 2.14.

[7] Excluding the two extremes where we calculate the loss for a single example only, or for all the training examples at once.

small predetermined positive value η, called the learning rate, and subtract the resulting term from the current values of the parameters.

We can therefore express the update:

$$(W, b) \leftarrow (W, b) - \frac{\eta}{|\mathcal{B}|} \sum_{i \in \mathcal{B}_t} \partial_{(W,b)} l_i(W, b). \tag{2.6}$$

2.2.3 Key concepts of deep learning

We described neural networks in subsection 2.2.2 as linear transformations with an added bias. A neural network only maps inputs directly to outputs via a single linear transformation, followed by an optional non-linear function. If our labels were truly related to the input data by a simple affine transformation, this approach would be sufficient. However, this is not always the case [55].

2.2.3.1 Incorporating hidden layers

We can overcome the limitations of single-layer models by incorporating one or more hidden layers. The simplest way to do this is to stack L fully connected layers on top of each other. Each layer feeds into the layer above it, until we generate outputs. We can think of the first $L - 1$ layers as our representation and the last layer as our linear predictor. This architecture is commonly called a multilayer perceptron, often abbreviated as MLP [54, 55].

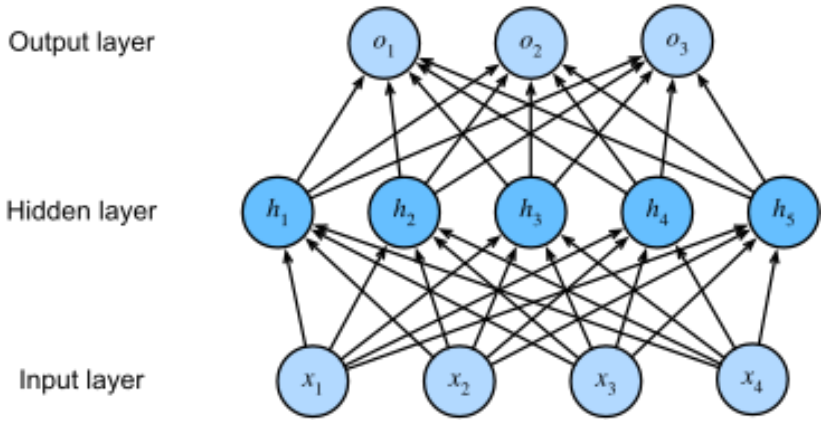


FIGURE 2.7: Architecture of a multilayer perceptron

The MLP illustrated in Figure 2.7 has four inputs, three outputs, and its hidden layer contains five hidden units. Since the input layer does not involve any computation, producing outputs with this network requires implementing the computations for the hidden and output layers; thus, the number of layers in this MLP is two. Note that both layers are fully connected. Each input influences every neuron in the hidden layer, and each of these in turn influences every neuron in the output layer [54, 55].

2.2.3.2 Mathematical formalization of a deep neural network

We denote by the matrix $X \in \mathbb{R}^{n \times d}$ a minibatch of n examples where each example has d inputs (features). For an MLP with a single hidden layer that has h hidden units, we denote by $\mathbf{H} \in \mathbb{R}^{n \times h}$ the outputs of the hidden layer, which are hidden representations. Since both the hidden and output layers are fully connected, we have hidden layer weights $W^{(1)} \in \mathbb{R}^{h \times d}$ and biases $b^{(1)} \in \mathbb{R}^h$, as well as output layer weights $W^{(2)} \in \mathbb{R}^{q \times h}$ and biases $b^{(2)} \in \mathbb{R}^q$. This allows us to compute the outputs $\mathbf{O} \in \mathbb{R}^{n \times q}$ of the single-hidden-layer MLP as follows:

$$\mathbf{H} = \sigma(XW^{(1)\top} + b^{(1)}), \tag{2.7}$$

$$\mathbf{O} = \mathbf{H}W^{(2)\top} + b^{(2)}. \tag{2.8}$$

To build more general MLPs, we can continue to stack such hidden layers, for example $\mathbf{H}^{(1)} = \sigma^{(1)}(XW^{(1)\top} + b^{(1)})$ and $\mathbf{H}^{(2)} = \sigma^{(2)}(\mathbf{H}^{(1)}W^{(2)\top} + b^{(2)})$, on top of each other, resulting in increasingly expressive models.

Optimization proceeds by following the same concept explained in paragraph 2.2.2.3. By applying the chain rule,

$$\frac{\partial z}{\partial x} = \frac{\partial z}{\partial y} \cdot \frac{\partial y}{\partial x}, \tag{2.9}$$

we can compute the gradient of the loss function with respect to weights and biases at each layer, starting from the output layer and working backwards to the input layer, which is why this operation is called backpropagation [54].

2.2.3.3 Activation functions

The outputs of the σ functions are called *activations*. Activation functions decide whether a neuron should be activated or not, they are differentiable operators that transform input signals into outputs, most of them add non-linearity [55].

- **Rectified Linear Unit (**ReLU**):** the most popular choice, due to both simplicity of implementation and good performance on a variety of predictive tasks. ReLU [58] provides a very simple non-linear transformation. Given an element x, the function is defined as the maximum of that element and 0:

$$\mathrm{ReLU}(x) = \max(x, 0). \tag{2.10}$$

- **Sigmoid function (**sigmoid**):** the sigmoid function maps inputs whose values lie in the domain $\mathbb{R}$ to outputs that lie in the interval $]0, 1[$. For this reason, the sigmoid function is often called a squashing function: it squashes any input in the range $]-\infty, +\infty[$ to a value in the range $]0, 1[$:

$$\mathrm{sigmoid}(x) = \frac{1}{1 + \exp(-x)}. \tag{2.11}$$

The sigmoid function is still widely used as an activation function on output units when we want to interpret the outputs as probabilities for binary classification problems [54]. However, it has largely been superseded by the simpler and more easily trainable ReLU function for most uses in hidden layers.

- **Hyperbolic tangent (**tanh**):** like the sigmoid function, the tanh function also squashes its inputs, transforming them into elements of the interval between -1 and 1:

$$\tanh(x) = \frac{1 - \exp(-2x)}{1 + \exp(-2x)}. \tag{2.12}$$

Although the shape of the function is similar to that of the sigmoid function (see Figure 2.8), the tanh function exhibits point symmetry around the origin of the coordinate system [59].

- **Softmax function:** the softmax function is commonly used to transform raw scores into probabilities. It is particularly well-suited for classification tasks. Suppose we have a vector of scores $\mathbf{z} = (z_1, z_2, \ldots, z_k)$, where each z_i is the score associated with class i. The softmax function transforms these scores into normalized probabilities

$$\hat{y}_i = \frac{\exp(z_i)}{\sum_j \exp(z_j)}, \tag{2.13}$$

where $\hat{y}_i$ is the probability that the input belongs to class i. The sum of $\hat{y}_i$ for all classes is guaranteed to be equal to 1, which is essential for probabilistic interpretation [55]. The softmax function is often used in conjunction with the cross-entropy loss function (see item 2.2.3.4).

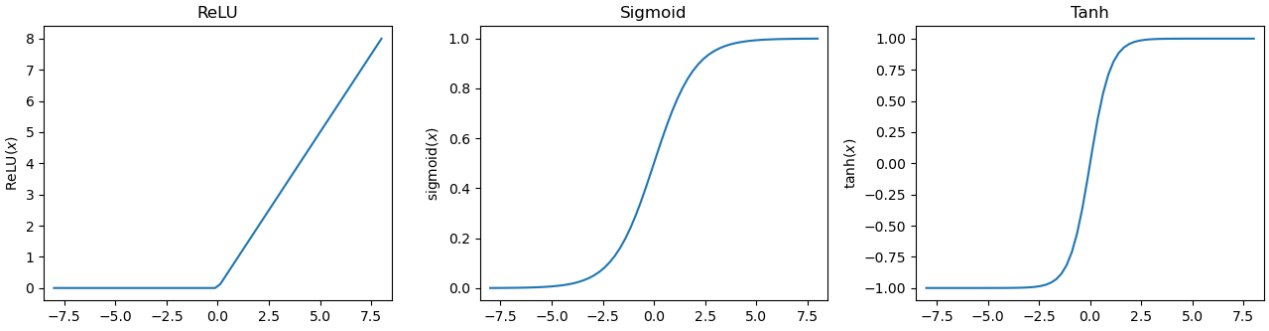


FIGURE 2.8: Curves of the mentioned activation functions

Note, however, that research on activation functions has not stopped. For example, the GELU (Gaussian Error Linear Unit) activation function $\text{GELU}(x) = x\Phi(x)$[8] by Hendrycks and Gimpel [60] and the $\text{Swish}(x) = x\sigma(x\beta)$[9] activation function proposed by Ramachandran et al. [61] can provide better accuracy in many cases.

[8] $\Phi(x)$ is the standard Gaussian cumulative distribution function.
[9] $\sigma(x)$ refers to the sigmoid function.

2.2.3.4 Learning and optimization

Although optimization provides a means to minimize the loss function for deep learning, in essence, the objectives of optimization and deep learning are fundamentally different. The former is primarily concerned with minimizing an objective while the latter aims to find a suitable model, given a finite amount of data [54]. This is what motivates the research and design of better optimization algorithms.

Loss functions

- **Loss functions for regression:**
  - *Mean Squared Error (MSE):* the mean squared error, mentioned earlier, is used to evaluate the quality of a regression model. It measures the average of the squares of the differences between predicted values and actual values [54, 55]:

$$\text{MSE}(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \tag{2.14}$$

  - *Mean Absolute Error (MAE):* is another similar error measure for regression models. It calculates the average of the absolute values of the differences between predicted values and actual values [54, 55]:

$$\text{MAE}(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|. \tag{2.15}$$

- **Cross-entropy loss:** is used to evaluate the quality of classification models. It measures the divergence between the probability distributions of actual and predicted classes [55]:

$$\text{CrossEntropyLoss}(y, \hat{y}) = -\sum_{i=1}^{n} y_i \log(\hat{y}_i) - (1 - y_i)\log(1 - \hat{y}_i). \tag{2.16}$$

- **Loss functions for separation[10]:** the hinge loss is commonly used in separation models (e.g., SVMs). It measures the margin between predicted scores and true labels [43]:

$$\text{HingeLoss}(y, \hat{y}) = \max(0, 1 - y \cdot \hat{y}). \tag{2.17}$$

Optimization algorithms We will present some optimization algorithms without going into the details of their inner workings, as that is beyond our main objective.

- **Adaptive Moment Estimation (Adam):** Adam integrates several effective techniques for deep learning optimization, offering robustness and efficiency. Introduced by Kingma and Ba in

[10] Separation is a form of binary classification.

2014 [62], Adam addresses issues such as handling redundant data and improving convergence speed. Despite its popularity, it can encounter divergence problems, which have been addressed by subsequent algorithms like Yogi proposed by Zaheer et al. in 2018 [63].

- **Root Mean Square Propagation (RMSProp):** is an optimization algorithm that uses a moving average of squared gradients to adjust the learning rate [64].

It is useful to mention that changing the learning rate during training can lead to improved accuracy and reduced overfitting[11] of the model [54].

A piecewise decrease of the learning rate whenever progress has stalled is effective in practice, it ensures that we converge efficiently to an appropriate solution, then only reduce the inherent variance of the parameters by decreasing the learning rate [62, 65].

2.2.4 Attention mechanisms

The evolution of deep learning has seen many innovations, but the emergence of attention mechanisms marks a significant turning point. Much like humans, the attention mechanism enables models to focus on specific parts of the input which facilitates the discernment of patterns and interconnections within the data. This emerges from the idea that some parts of the input are assumed to be more relevant than others. Attention is what allows models to quantify the degree of importance each part should receive [66].

Attention mechanisms were initially introduced to enhance encoder-decoder recurrent neural networks (RNNs) for sequence-to-sequence tasks, such as machine translation. The idea was proposed by Bahdanau et al. [67] as a way to allow the decoder to focus on different parts of the input sequence at each step of the output generation, rather than relying on a single fixed-length vector representation of the input. Attention mechanisms not only improved the performance of RNNs in machine translation but also provided insights into the translation process. This ability to align and focus on relevant parts of the input sequence led to claims of increased interpretability, although the precise meaning and interpretation of attention weights remain an area of active research [68].

The significance of attention mechanisms grew with the introduction of the Transformer model by Vaswani et al. [69]. The Transformer architecture eliminates recurrent connections altogether and relies solely on self-attention mechanisms to model relationships between all input and output tokens. This innovation has dramatically improved the performance of models in natural language processing tasks. Transformers have become the backbone of state-of-the-art models for a wide range of applications, from simple regression tasks to language understanding, computer vision, and other sub-fields of deep learning.

Attention applies linear transformations to the input to map the features into three distinct representations: queries Q, keys K, and values V. For a database $\mathcal{D}$ of m key-value pairs,

$$\mathcal{D} = \Big\{(\mathbf{k}_1, \mathbf{v}_1), \ldots, (\mathbf{k}_m, \mathbf{v}_m)\Big\}.$$

[11] Overfitting occurs when the model fits too closely to the training data, thereby losing its ability to generalize effectively on new data.

Denoted by $\mathbf{q}$ a query, the attention over $\mathcal{D}$ can be defined as such

$$\text{Attention}(\mathbf{q}, \mathcal{D}) = \sum_{i=1}^{m} \alpha(\mathbf{q}, \mathbf{k}_i)\mathbf{v}_i, \tag{2.18}$$

where $\alpha(\mathbf{q}, \mathbf{k}_i) \in \mathbb{R}, i = 1, \ldots, m$ are the attention weights.

They are typically computed using a similarity measure (score function) a between the query $\mathbf{q}$ and the keys $\mathbf{k}_i$, then into a distribution function of choice, commonly, the softmax function,

$$\alpha(\mathbf{q}, \mathbf{k}_i) = \frac{\exp(a(\mathbf{q}, \mathbf{k}_i))}{\sum_j \exp(a(\mathbf{q}, \mathbf{k}_j))}. \tag{2.19}$$

The attention mechanism gives a differentiable way of control that allows a neural network to select elements from a set and generate an associated weighted sum over representations [70].

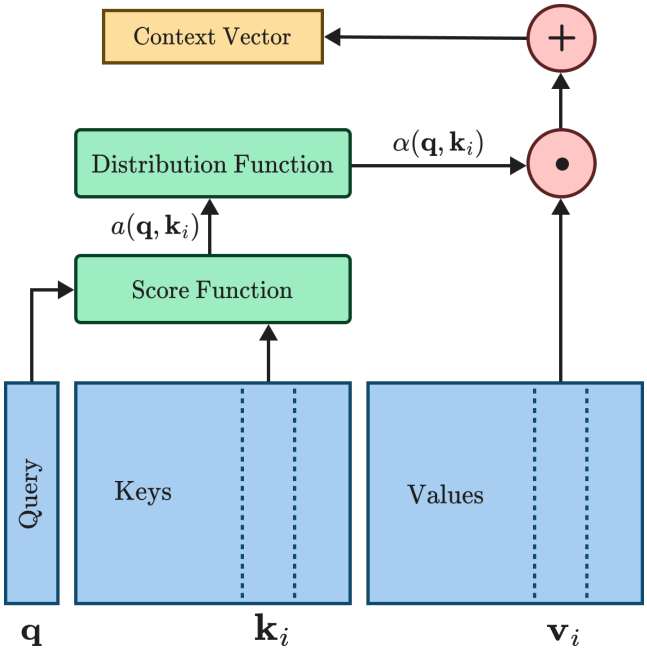


FIGURE 2.9: Schematic diagram of a typical attention mechanism

2.2.4.1 Self-attention

Self-attention, also known as intra-attention, is a specific type of attention mechanism where the queries, keys, and values are all derived from the same source. For an input sequence each token is transformed into a query, key, and value vector. The attention score is then processed in the same way as general attention mechanisms, resulting in a new representation for each token that incorporates contextual information from the entire sequence [68].

2.2.5 Ensemble learning

Ensemble learning offers a reliable strategy to improve predictive models and has become a prominent paradigm in machine learning. It leverages the strengths of several base learners while mitigating individual weaknesses [71]. This aggregation of diverse models facilitates a deeper understanding of underlying data patterns, which leads to enhanced prediction accuracy and better generalization capabilities. The motivation behind ensemble learning is to address the limitations of individual models, such as high variance, bias, or sensitivity to noise. Across various applications, ensemble techniques consistently outperform standalone models, making them a valuable asset for complex machine learning tasks.

2.2.5.1 Types of ensemble methods

Ensemble learning methods can be broadly categorized into three main types: bagging, boosting, and stacking.

- **Bagging:** bagging, or bootstrap aggregating, is a technique designed to mitigate model variance. It achieves this by training multiple models on different subsets of the data and then combining their predictions. A prominent example of bagging is the Random Forest algorithm, proposed by Breiman [72]. Random Forest constructs an ensemble of decision trees during training and produces the mode of class labels or the mean prediction from the individual trees. This approach effectively addresses overfitting and enhances generalization performance.
- **Boosting:** boosting aims to enhance weak learners by training models sequentially. Each subsequent model corrects the errors made by its predecessor. A notable boosting algorithm is AdaBoost, proposed by Freund and Schapire [73]. AdaBoost assigns weights to each data instance, adjusting them based on the previous model's performance. This iterative process emphasizes challenging instances, ultimately improving overall model accuracy. Other well-known variants of boosting include Gradient Boosting Machines (GBM) and XGBoost.
- **Stacking:** stacking, also known as stacked generalization, constitutes an ensemble technique that integrates multiple models through a meta-learner. In this methodology, base models are trained on the dataset, and their predictions serve as input features for a higher-level meta-model. The primary objective of the meta-model is to capture the relationships between the predictions generated by the base models and the target variable. Wolpert [74] introduced stacking, and demonstrated its efficacy in enhancing predictive performance by effectively amalgamating diverse model predictions.

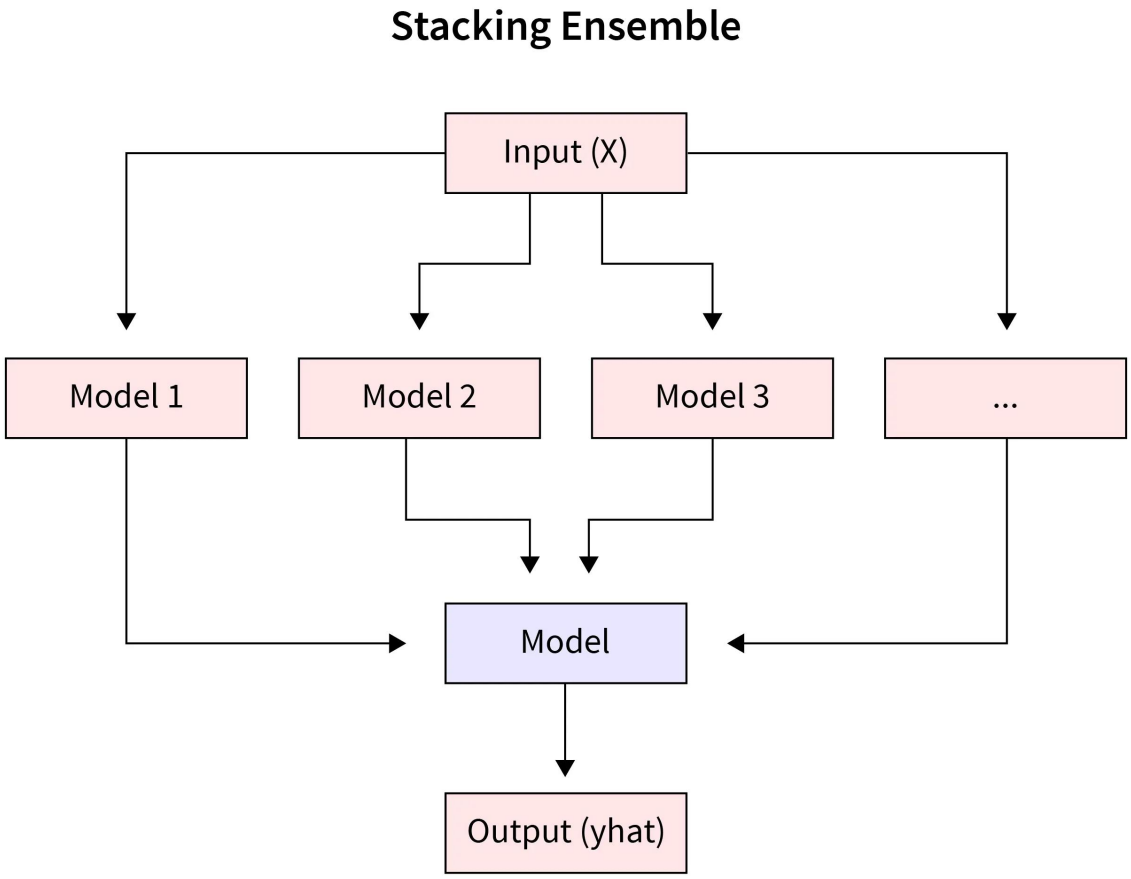


FIGURE 2.10: Architecture of a stacking model [75]

2.3 Conclusion

This chapter has provided a comprehensive overview of the fundamentals and techniques in machine learning and deep learning. We began by exploring the basic principles of machine learning, its categories, tasks, and limitations, setting the stage for a deeper dive into artificial neural networks. We then progressed to the key concepts of deep learning, including the incorporation of hidden layers, mathematical formalization of deep neural networks, activation functions, and optimization techniques. Finally, we introduced the attention mechanism and ensemble learning framework, both innovations that pave the way for the development of increasingly sophisticated algorithms that improve the performance of models for better accuracy and robustness.

Chapter 3

Photocatalytic degradation model

3.1 Introduction

In our pursuit of sustainable water treatment solutions, photocatalytic degradation has emerged as a promising technique. However, the complexity of experimentation in this field, with its numerous interdependent parameters, has hindered rapid progress. We glanced over how machine learning, particularly deep learning models, can accelerate progress in multiple subfields of chemistry.

In this chapter, we present a novel approach: a data augmenting self-attention network (DASAN) that predicts photocatalytic degradation rates and efficiencies from a set of experimental parameters. Our model is designed to extract valuable insights from limited datasets, a common challenge in this field. By combining expert-guided data augmentation with self-attention mechanisms. This work aims to guide researchers toward optimal conditions for their photocatalytic experiments.

3.2 Photocatalytic degradation

Photocatalytic degradation involves using a catalyst, typically a semiconductor like titanium dioxide (TiO_2), to speed up the breakdown of pollutants when exposed to light, usually in the ultraviolet range [76]. The catalyst generates electron-hole pairs that facilitate redox reactions, decomposing contaminants into less harmful substances such as carbon dioxide and water [77].

Photocatalytic degradation is an environmentally significant process that utilizes abundant sunlight energy to decompose a wide range of pollutants. Its key advantages include complete mineralization of pollutants, eliminating the need for water disposal, low operational costs, and the requirement of only mild temperature and pressure conditions [78]. By addressing pollution control and supporting renewable energy conversion, photocatalysis offers a sustainable alternative. Additionally, it effectively tackles the global rise in organic contaminants, especially those resistant to traditional biological treatment methods. Harnessing solar energy, photocatalysis facilitates chemical reactions that transform toxic substances into less harmful forms, making it both economically viable and environmentally beneficial [17].

The efficiency of photocatalytic degradation processes is influenced by multiple variables related to the photocatalyst material itself, the properties of the target pollutant, and the environmental

conditions under which the degradation takes place. These factors must be carefully considered and optimized in order to maximize degradation rates and treatment efficacy [79].

3.2.1 Mechanism

The photocatalytic degradation process involves several key steps that occur when a semiconductor photocatalyst like TiO_2 is irradiated with light of sufficient energy. The general mechanism [17] can be summarized as follows:

1. **Photoexcitation:** when a photocatalyst absorbs photons with energy that is equal or exceeds its band gap, electrons are elevated from the valence band to the conduction band, thus electron-hole pairs are formed.
2. **Charge carrier separation:** the photogenerated electrons and holes migrate to the surface of the photocatalyst particle.
3. **Redox reactions:** the separated charge carriers participate in redox reactions, where electrons reduce dissolved oxygen (O_2) to form superoxide radicals $(O_2)^-$, while the holes oxidize water (H_2O) or hydroxideions (OH^-) to generate hydroxyl radicals (OH).
4. **Pollutant degradation:** the highly reactive OH and $(O_2)^-$ species attack organic pollutants adsorbed on the catalyst surface or in the surrounding medium. These reactions lead to mineralization into simpler compounds (e.g., CO_2, H_2O, and inorganic ions).

For TiO_2 (anatase phase), which possesses a band gap of approximately 3.2 eV, ultraviolet light initiates photoexcitation. The relevant reactions are as follows:

$$\mathrm{TiO_2} + h\nu \rightarrow e^- + h^+ \tag{3.1}$$

$$e^- + \mathrm{O_2} \rightarrow \mathrm{O_2^-} \tag{3.2}$$

$$h^+ + \mathrm{H_2O} \rightarrow \mathrm{OH} + \mathrm{H^+} \tag{3.3}$$

The hydroxyl radicals are highly oxidizing and can rapidly attack organic molecules, which leads to their mineralization into CO_2, H_2O and inorganic ions. The superoxide radicals can also participate in the degradation process or form additional OH radicals through further reactions.

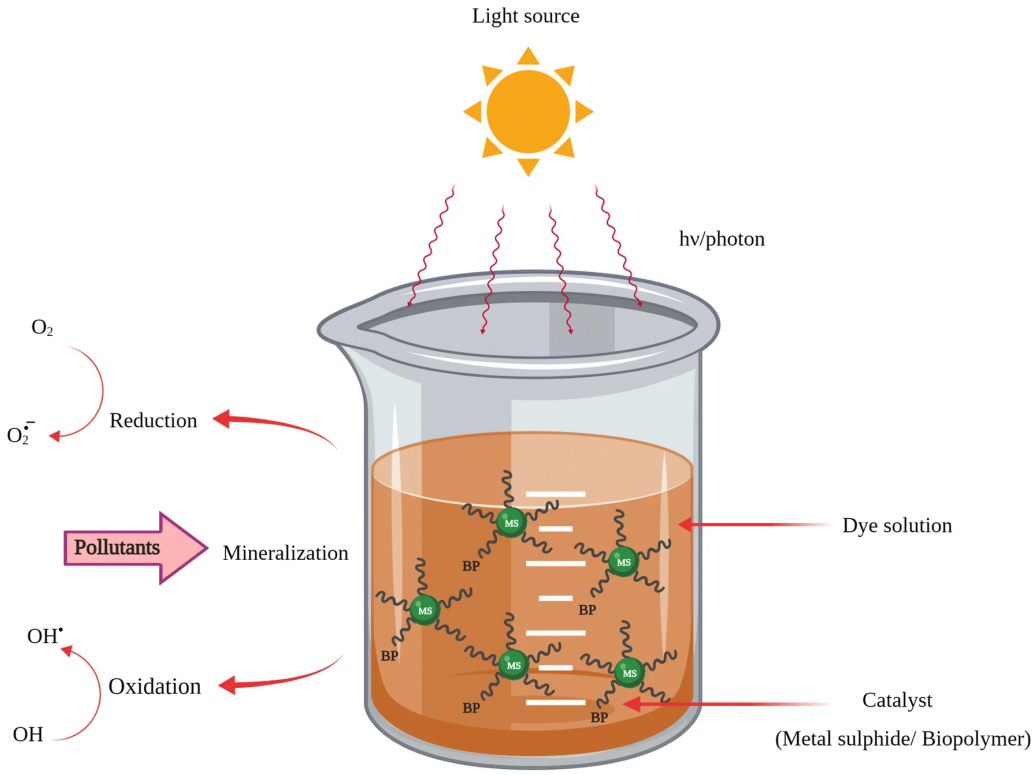


FIGURE 3.1: Illustration of the Photocatalytic Degradation Process [80]

3.2.2 Photocatalyst properties

1. **Specific surface area and structure:** catalysts with high surface areas and optimized nanostructures provide more reaction sites, facilitating improved adsorption and subsequent degradation of pollutants. Tailoring the catalyst morphology is thus an important design consideration.

2. **Dopants and modifications:** introducing dopant species like metals or non-metals into photocatalysts can enhance light absorption, inhibit electron-hole recombination, and generate more oxidizing radicals - all boosting degradation capability.

3. **Crystal phase:** the crystal structure of the photocatalyst influences its electronic behavior, charge transport properties, and surface chemistry, parameters that directly impact degradation performance.

3.2.3 Pollutant characteristics

1. **Initial concentration:** higher initial pollutant concentrations generally reduce degradation efficiency due to fewer active sites being available and screening effects from the pollutant molecules themselves.

2. **Molecular structure:** pollutant molecules with electron-withdrawing functional groups tend to adsorb more readily and degrade faster compared to those with electron-donating groups.

3. **Reactivity and byproducts:** some pollutants and their degradation intermediates are inherently more recalcitrant, hindering complete mineralization.

3.2.4 Environmental factors

1. **Solution pH:** this variable affects the surface charge of the catalyst and the ionization state of pollutants, thereby altering adsorption interactions that initiate degradation.
2. **Temperature:** moderate heating can promote pollutant adsorption and reactive species mobility to accelerate degradation, but excessively high temperatures deactivate catalysts.
3. **Light properties:** higher light intensities provide more photons to drive reactions, while the wavelength must match the catalyst's absorption for electron-hole generation.
4. **Dissolved oxygen:** oxygen acts as an electron acceptor, that inhibit charge recombination and facilitate oxidation reactions critical for degradation.
5. **Inorganic ions:** certain ionic species can competitively adsorb, quench radicals, or promote charge carrier recombination - which may potentially enhance or suppress degradation.

3.2.5 Experimental methods

Photocatalytic degradation processes are typically studied using two primary experimental methods: batch reactor experiments and flow reactor systems. These methods are critical for producing the empirical data that underpins the computational models and simulations.

3.2.5.1 Flow reactor systems

Flow reactor systems offer an alternative method for studying photocatalytic degradation under dynamic conditions. These systems consist of a transparent reaction chamber through which the reaction mixture flows continuously. The flowing mixture is exposed to light irradiation, triggering the photocatalytic reaction. Flow reactors offer several advantages over batch reactors. They allow for better control over reaction parameters such as residence time, flow rate, and light intensity. Additionally, they are well-suited for kinetic studies and facilitate the scale-up of photocatalytic processes from the laboratory to industrial scales [81].

3.2.5.2 Batch reactor experiments

Batch reactor experiments are a common method used to study photocatalytic degradation under controlled conditions. The setup for these experiments often includes a reaction vessel equipped with a light source to initiate the photocatalytic reaction. A magnetic stirrer ensures uniform mixing of the reaction mixture, which usually consists of a suspension of the photocatalyst and the pollutant. Monitoring devices are used to track reaction parameters such as temperature and pH.

In these kind of experiments, the reaction mixture is irradiated with light of the appropriate wavelength, and the degradation kinetics of the pollutant are monitored over time. This is usually done by sampling the reaction mixture at regular intervals and analyzing the concentration of the pollutant [82].

3.3 Related work

Recent research has witnessed a surge in the development of ANN models aimed at predicting degradation rates and efficiencies of various pollutants under diverse experimental conditions [83].

Jiang et al. [84] presented an ANN model that estimates the rate constants of 76 organic water contaminants with TiO_2 as the photocatalyst under UV irradiation. It considers six experimental variables: UV intensity, TiO_2 dosage, contaminant concentration, initial pH, temperature, and contaminant type. The model demonstrates good accuracy with a root mean square error (RMSE) of 0.173 [84].

Garg et al. [85] used a commercial TiO_2 catalyst to photodegrade the aqueous dye Acid Red 114 (AR114) under UV light. The response surface method (RSM) was used to enhance the photocatalytic degradation process. The experimental input parameters included TiO_2 dose, solution pH, initial AR114 concentration, time and area/volume, and UV light intensity. The output was AR114's degradation and decolorization efficiency. Within 150 minutes of light irradiation, the degradation efficiency reached 100%. The proposed 6:7:2:2 feedforward model, with a relative correlation coefficient of 0.998, provides a very good prediction of experimental data [85].

Chandrika et al. [86] proposed an ANN model to predict the photocatalytic degradation of Malachite Green (MG) dye with TiO_2 photocatalyst. The TiO_2 was created using the sol-gel synthesis process. The optimum condition was determined by varying the experimental parameters, the photocatalytic degradation efficiency was 90% at the optimum conditions of 3 mg TiO_2 nanoparticle, 60 minutes of reaction time, and 20 ppm initial dye concentration. The back-propagation algorithm was used for ANN modeling, and the best efficiency was predicted with a 3-6-1 topology and an R^2 value of 0.9707 [86].

Boutra et al. [87] suggested an ANN model with a Bayesian regularization algorithm for photocatalytic degradation of solophenyl brown AGL dye and paracetamol using a titanium dioxide photocatalyst exposed to sunlight. RSM was used to optimize the photocatalytic experiments, which involved varying experimental parameters such as titanium dioxide dose, solution pH, and pollutant concentration. Optimal conditions of 0.7 g/L catalyst loading, 0.34 initial concentration ratio of pollutants, and pH 6.5 solution resulted in a degradation rate of approximately 99%. The ANN model with the optimal topology 3-4-3 yielded a correlation coefficient of 0.99 for co-degradation yield, indicating excellent agreement between predicted and observed data [87].

3.4 Dataset

The data in this work originates from Bouazza's work [26] on the photocatalytic degradation of Methylene Blue (MB) using TiO_2/curcumin nanocomposite material [26]. The experiments were conducted using a custom-designed continuous-flow reactor system under natural pH and at a temperature of 25°C. A TiO_2/5% curcumin nanocomposite material was deposited on a cellulose paper substrate via a dip-coating method, which was then placed inside a quartz tube reactor, surrounded by UV lamps to provide the irradiation source. A pump enabled the continuous flow of the aqueous methylene blue solution through the reactor. Three primary parameters: mass of TiO_2/5% curcumin material (mg), light intensity (w/cm^2), and initial MB concentration (ppm), were isolated to study their influence on the degradation's performance. Samples were collected at regular intervals during the photocatalytic runs and analyzed by UV-Vis spectroscopy to monitor the degradation of MB. The degradation followed a first-order kinetic model, described by the equation:

$$[\text{MB}] = Ae^{-kt} + E \tag{3.4}$$

where A is the amplitude, k is the first-order rate constant, t is time, and E is the endpoint.

Although three variations were done on each studied parameter, the work encompasses only seven distinct experiments instead of the anticipated nine. This is because three identified optimum values were used repeatedly across the experimentation [26].

3.5 Data preprocessing

Depending on the behavior of the degradation or the availability of the data, experimental metrics, parameters of various curve functions, or simply y-values of the degradation can be set as targets. The resulting dataset $\mathcal{D}$ of k rows is defined as a set of tuples of n feature columns and m target columns,

$$\mathcal{D} = \left\{ (x_i^1, x_i^2, \ldots, x_i^n) \frown (y_i^1, y_i^2, \ldots, y_i^m) \mid 1 \leq i \leq k \right\}.$$

Bouazza et al [26] observed that the degradation of MB adheres to a first-order kinetic model [26]. This degradation can be equivalently represented by the following exponential decay function:

$$f(t) = (N_0 - c)e^{-\lambda t} + c. \tag{3.5}$$

In this equation, N_0 represents the initial value, λ is the decay rate, and c is the asymptotic value as time progresses. This representation effectively reduces the number of parameters our model needs to estimate from three to two. The eliminated parameter being the amplitude A, which is dependent on a known constant N_0.

To better fit our experimental data, we adjusted the initial value N_0 to the experimental average of ~ 0.96, instead of the theoretical 1.0. Due to experimental uncertainty, the disparity between the residual concentration and the fitted asymptotic value c is not of concern.

For each experimental degradation, we calculate a vector of y-values by applying the fitted exponential decay function to each element in the time vector of size 899,

$$\vec{y} = \left[f(t_1), f(t_2), ..., f(t_{899})\right].$$

The resulting vector (see figures 3.2, 3.3, 3.4) is assigned as the target to the combination of its respective experimental parameters: mass of TiO_2/5% curcumin material (mg), light intensity (w/cm^2), and initial MB concentration (ppm).

$$\mathcal{D} = \left\{(p_i^1, p_i^2, p_i^3) \frown (\vec{y}_i) \mid 1 \leq i \leq k\right\}. \tag{3.6}$$

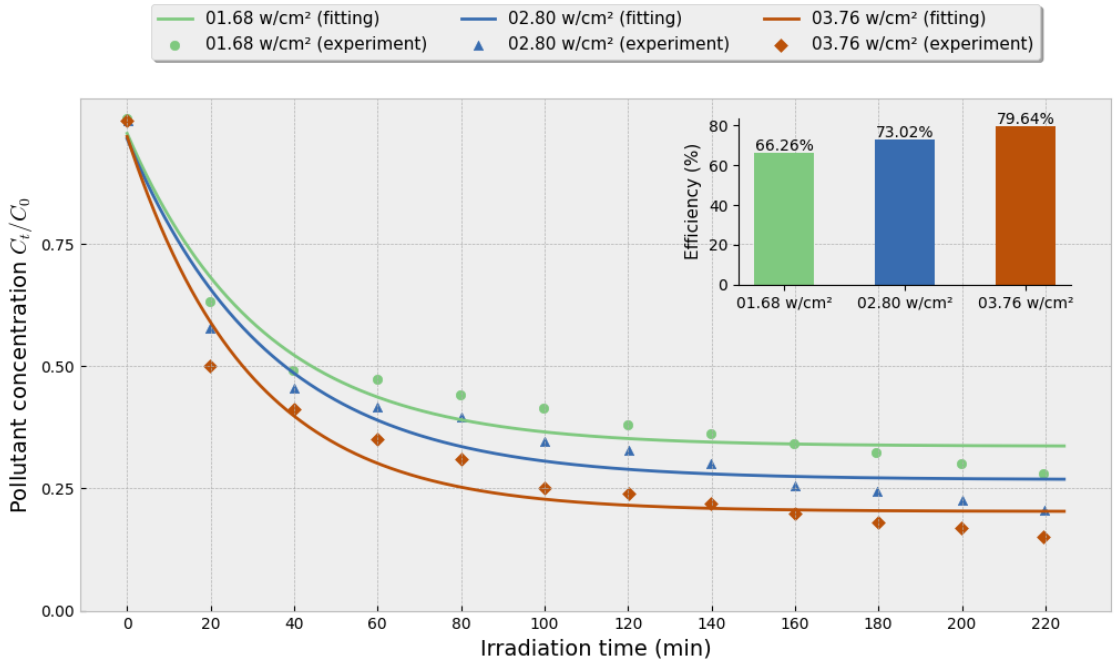


FIGURE 3.2: Effects of light intensity variation over the degradation under the control group: material mass = 14 mg, initial MB concentration = 10 ppm

3.6 Data partitioning

Building upon Bouazza et al [26] works's efforts on studying the effects of each experimental parameter on the performance of the degradation [26]. Our dataset $\mathcal{D}$ is partitioned into multiple subsets I_u^j, each representing the variation of a single experimental parameter p^j conditioned on a distinct combination u of control variables, which in our case are the other two parameters. Even though this does not apply to our data at hand, this means that multiple subsets may be created for one given experimental parameter, each corresponding to a unique set of values for the control group. This technique relies heavily on the experimental methodology, which is more elaborated on in section 3.10.2.

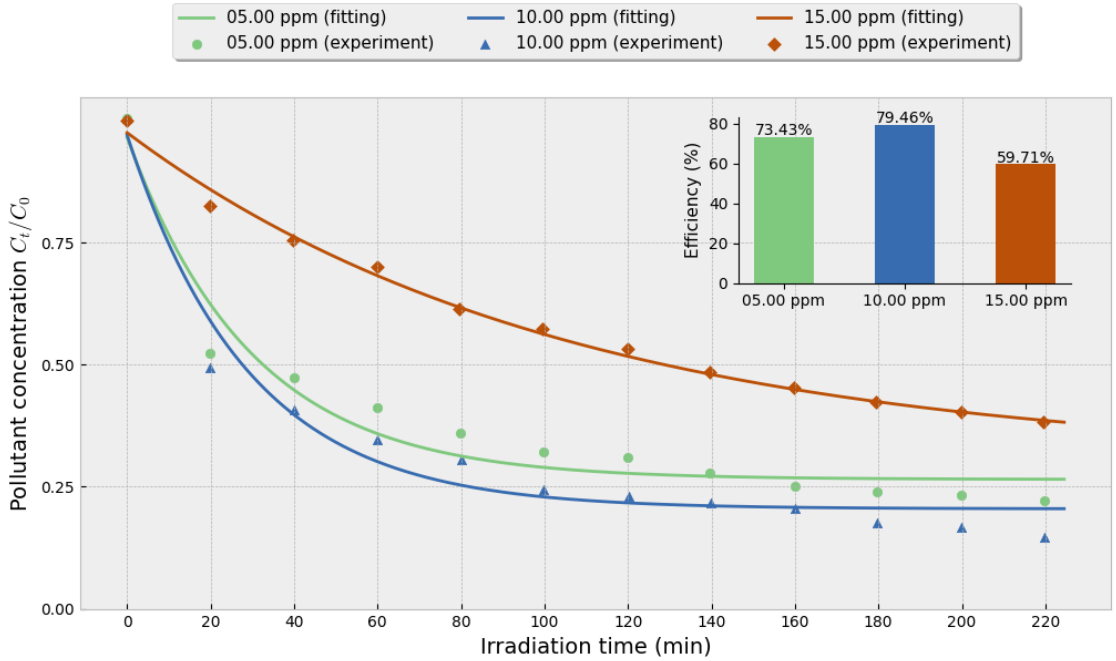


FIGURE 3.3: Effects of initial pollutant concentration variation over the degradation under the control group: material mass = 14 mg, light intensity = 3.76 w/cm²

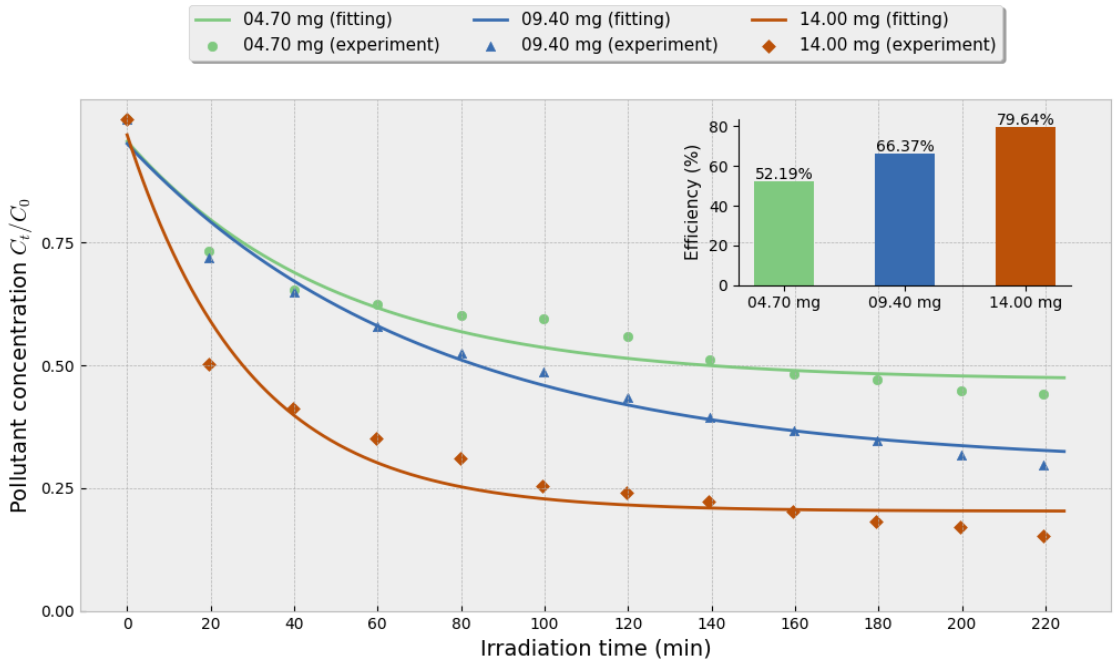


FIGURE 3.4: Effects of material mass variation over the degradation under the control group: light intensity = 3.76 w/cm², initial pollutant concentration = 10 ppm

3.7 Base model

A homogeneous set of base models B_u^j is trained on each subset I_u^j to generate new training material. The aim is to model the localized behavior of the target degradation curve in relation to the change of the experimental parameter while implicitly accounting for the effects of the control group through the partitioning of the data.

The models' architecture comprises a DNN with an input layer of one unit followed by three hidden layers into an output layer of two units corresponding to scalar values for λ and c, denoted by $\tilde{\lambda}$ and $\tilde{c}$ with λ ranging from 0 to 0.1, and c ranging from 0 to N_0. The output of the DNN is fed into an exponential decay layer that scales the parameters appropriately and outputs $\vec{y}$. The model, when trained indirectly on the y-values rather than directly on the parameters of the function, was found to yield better results. More on that in section 3.10.1.

The beta model's networks and propagation in listings 3.1), 3.2.

```
def build(self, input_shape):
    self.dense_1 = Dense(self.TOPOLOGY[0], self.HIDDEN_ACTIVATION)
    self.dense_2 = Dense(self.TOPOLOGY[1], self.HIDDEN_ACTIVATION)
    self.dense_3 = Dense(self.TOPOLOGY[2], self.HIDDEN_ACTIVATION)
    self.dense_output = Dense(self.NUM_FUNC_PARAMS, self.OUTPUT_ACTIVATION)
```

LISTING 3.1: Base model: TensorFlow implementation of the `call()` method

```
def call(self, inputs):
    self.last_params = self.dense_output(self.dense_3(self.dense_2(self.dense_1(inputs))))
    return self.exp_decay(self.last_params)
```

LISTING 3.2: Base model: TensorFlow implementation of the `call()` method

Each base model B_u^j outputs its prediction over a predefined linear space s^j specific to p^j. Let n_preds be the number of predictions,

$$s_i^j = \frac{s_{\max}^j - s_{\min}^j}{\text{n\_preds}} i + s_{\min}^j, \tag{3.7}$$

for $i \in [1, n-1]$. $s_{\max}^j$ and $s_{\min}^j$ are the upper and lower bounds of the linear space respectively. Adjusting the limits influences the extent of interpolation or extrapolation exhibited by the base models. An augmented subset $\mathcal{A}_u^j$ is formed using these predictions.

Each parameter p^j and its respective combination(s) u of control variables are joined as the features, with their predicted $\vec{y}$ as the target. The augmented subsets are then concatenated into the augmented dataset $\mathcal{A}'$ of l rows:

$$\mathcal{A}' = \left\{ (p_i^1, p_i^2, p_i^3) \frown (\vec{y}_i) \mid 1 \leq i \leq l \right\}. \tag{3.8}$$

$\mathcal{A}'$ is then used as training material for our meta model M.

3.8 Meta model

Initially, we compute the query vector $\mathbf{q}_i \in Q$, the key vector $\mathbf{k}_i \in K$ and the value vector $\mathbf{v}_i \in V$ for each feature p^j by multiplying the feature by the W_q^j, W_k^j and W_v^j weight matrices, respectively (Eq. 3.9-3.11). The weight matrices are of h units and are learned parameters.

$$\mathbf{q}_j = W_q^j p^j, \tag{3.9}$$

$$\mathbf{k}_j = W_k^j p^j, \tag{3.10}$$

$$\mathbf{v}_j = W_v^j p^j. \tag{3.11}$$

Representations of higher dimensionality provide enough *room* for the attention mechanism to compute similarity scores more effectively; the number of units h in the transformation layers must be accounted for and optimized.

The attention unit calculates the context vector for each query $\mathbf{q} \in Q$ with the dot product of $\mathbf{q}_i$ and $\mathbf{k}_i^\top$ as its alignment function $a(\mathbf{q}, \mathbf{k})$,

$$\tilde{C} = \Big[\text{Attention}\left(\mathbf{q}, \{(\mathbf{k}_1, \mathbf{v}_1), \ldots, (\mathbf{k}_n, \mathbf{v}_n)\}\right) \mid \mathbf{q} \in Q \Big],$$

or in simpler notation,

$$\tilde{C} = \text{AttentionUnit}(Q, K, V). \tag{3.12}$$

The resulting matrix $\tilde{C} \in \mathbb{R}^{n \times h}$ is then subjected to a parameter-wise sum reduction to obtain the context vector C. The context vector is then fed as a latent representation into a DNN of three layers with an output dimension of two corresponding to $\tilde{\lambda}$ and $\tilde{c}$. Finally, the exponential decay layer computes the y-values vector. The full DASAN model is illustrated in figure 3.5.

Using the TensorFlow library, the meta model uses the following networks (implemented in listing 3.3). For each iteration, the propagation is implemented in listing 3.4.

```
def build(self, input_shape):
    self.NUM_OF_CONDS = input_shape[1]

    self.query_denses = [Dense(self.H) for i in range(self.NUM_OF_CONDS)]
    self.key_denses = [Dense(self.H) for i in range(self.NUM_OF_CONDS)]
    self.value_denses = [Dense(self.H) for i in range(self.NUM_OF_CONDS)]
    self.attention = Attention()

    self.dense_1 = Dense(self.DNN_TOPOLOGY[0], self.DNN_ACTIVATION)
    self.dense_2 = Dense(self.DNN_TOPOLOGY[1], self.DNN_ACTIVATION)
    self.dense_output = Dense(self.NUM_FUNC_PARAMS, self.OUTPUT_ACTIVATION)
```

LISTING 3.3: Meta model: TensorFlow implementation of the `build()` method

```
def call(self, inputs):
    params = tf.unstack(inputs, axis=1)

    Q = [f(tf.expand_dims(p, 1)) for p, f in zip(params, self.query_denses)]
    K = [f(tf.expand_dims(p, 1)) for p, f in zip(params, self.key_denses)]
    V = [f(tf.expand_dims(p, 1)) for p, f in zip(params, self.value_denses)]

    Q, K, V = tf.stack(Q, 1), tf.stack(K, 1), tf.stack(V, 1)
    C_tilde = self.attention([Q, K, V])
    C = tf.reduce_sum(C_tilde, 1)

    self.exp_decay_params = self.dense_output(self.dense_2(self.dense_1(C)))

    return self.exp_decay(self.exp_decay_params)
```

LISTING 3.4: Meta model: TensorFlow implementation of the `call()` method

3.9 Training

The models were implemented and trained locally using the CPU build of Google's TensorFlow library version 2.12.0. Our machine had an AMD Ryzen 5 5600G processor @ 3.9 GHz. All the weights were Glorot-initialized and adjusted using the Adam optimization with a learning rate of 0.001, a β_1 of 0.9, and a β_2 of 0.999 using a batch size of one. The base models are trained for 5000 epochs to ensure the convergence of the MSE loss function while the meta model was kept for 1500 epochs.

3.10 Results and discussion

3.10.1 Point-wise prediction and parameter-wise prediction

Originally, our base and meta models were trained to directly predict the parameters λ and c of the exponential decay function. However, minimizing the difference between the predicted and actual parameters suffered from the exponential nature of the function, where minor errors in parameter estimation led to significant discrepancies in the output y-values. Instead, we trained the model indirectly by first predicting the parameters, then passing them through an exponential decay layer. The loss was calculated based on the difference between the predicted y-values and the target degradation y-values. This directly optimized the model for the final output, which is ultimately our primary concern.

3.10.2 Dependence on experimental methodology

A key limitation of the current approach arises from its dependence on the underlying experimental methodology used to obtain the initial dataset. The data augmentation via the base models relies on

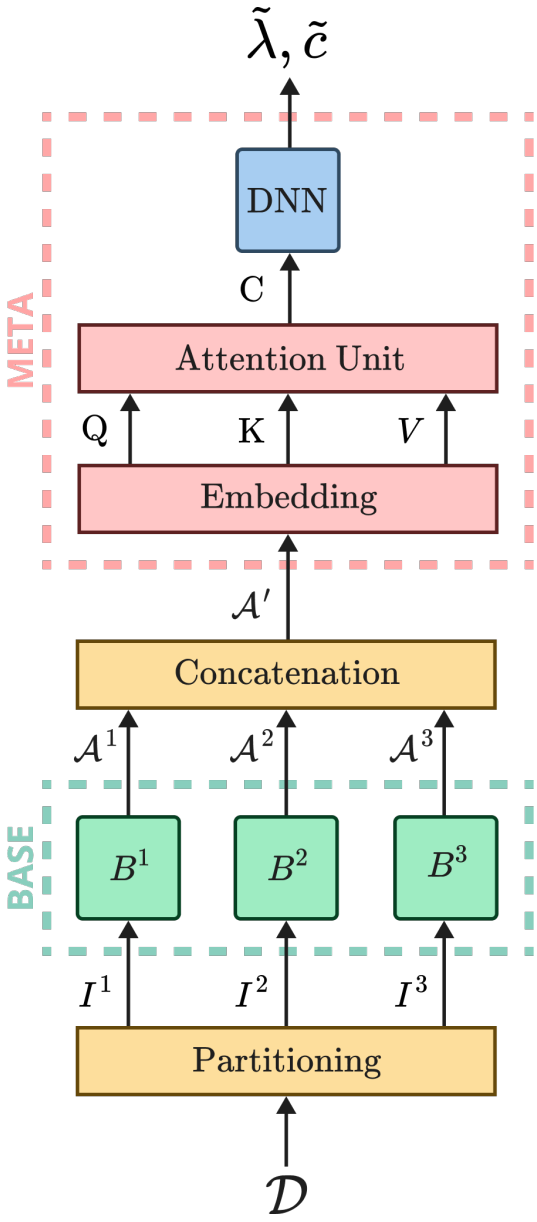
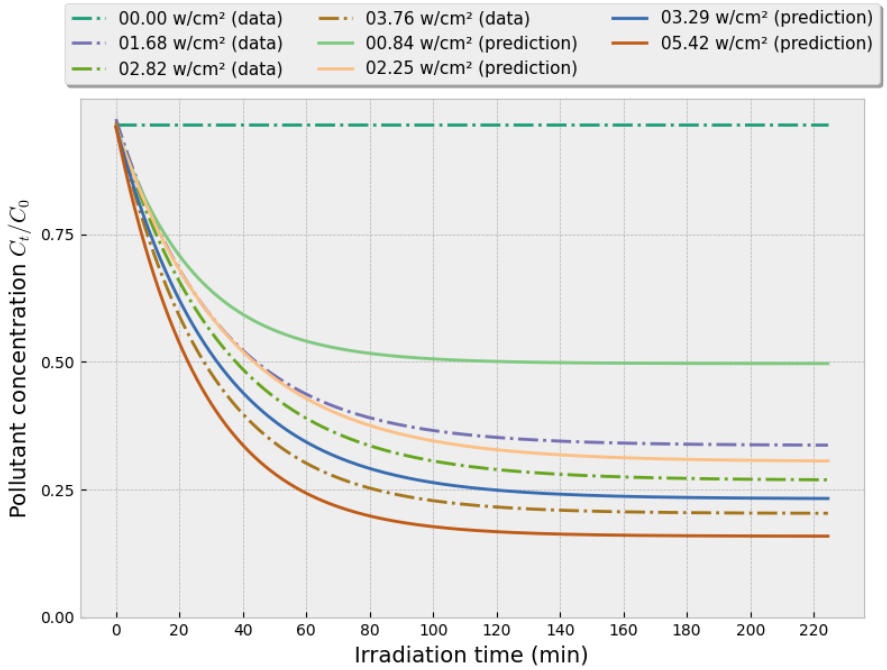
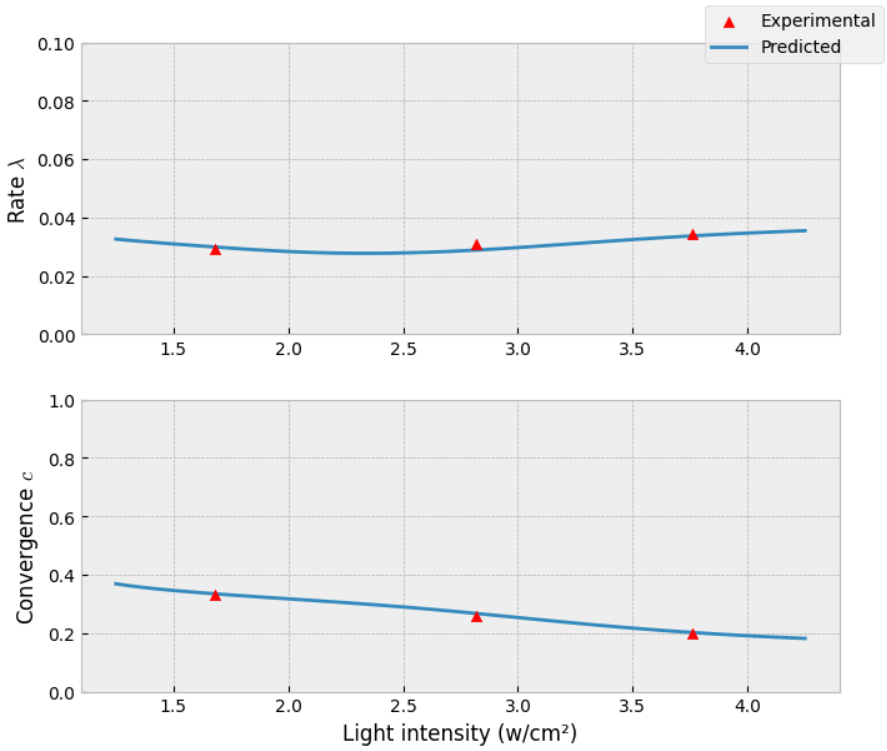


FIGURE 3.5: Schematic diagram of the end-to-end model architecture

varying each experimental parameter (photocatalyst mass, light intensity, initial pollutant concentration) individually. This one-factor-at-a-time experimental design enabled modeling the influence of each parameter but may fail to capture interdependencies or interaction effects that could arise when multiple parameters are varied simultaneously. While the data augmentation approach extracts more information from the limited dataset, it still carries the constraints imposed by the specific experimental design and methodological assumptions. To overcome this, a more comprehensive design involving the simultaneous variation of multiple factors would be required. However, this would substantially increase the experimental burden needed to sufficiently span the combined parameter space.

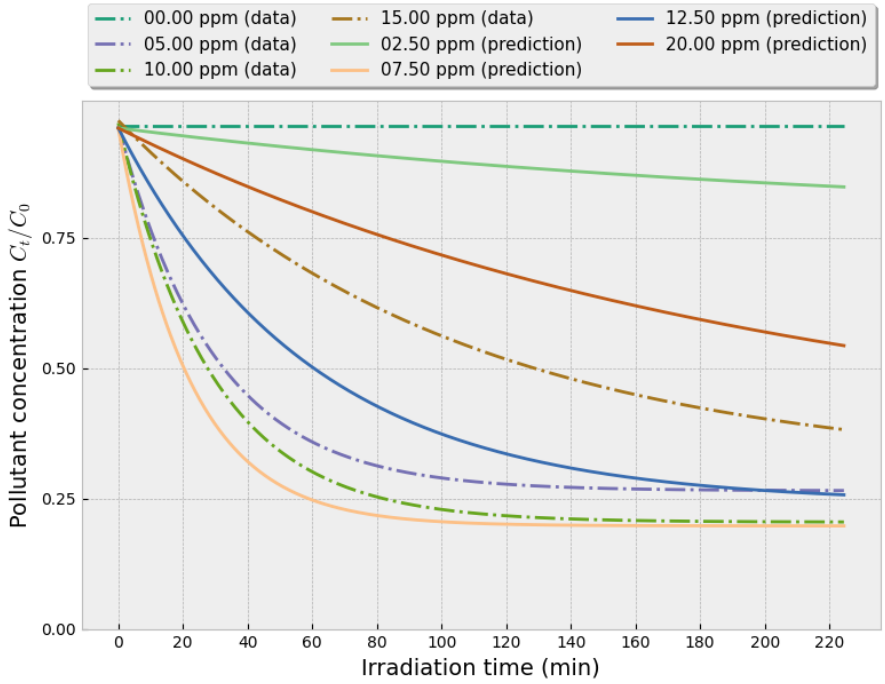


(a) Sample of predictions

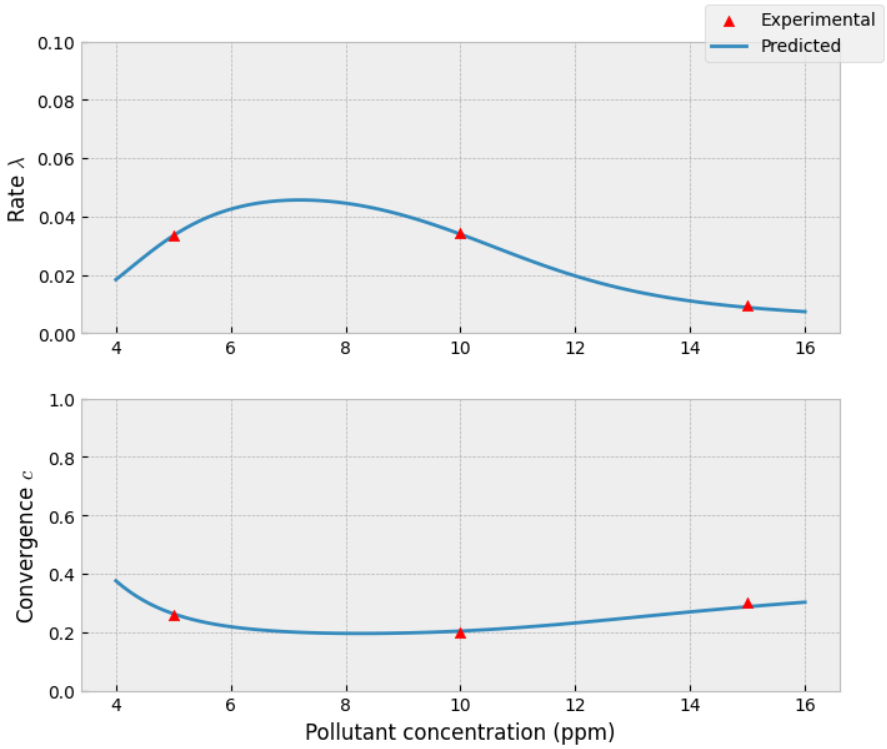


(b) Rate and convergence

FIGURE 3.6: Base model trained on the change of light intensity (TiO_2/curcumin material mass = 14 mg, initial MB concentration = 10 ppm)

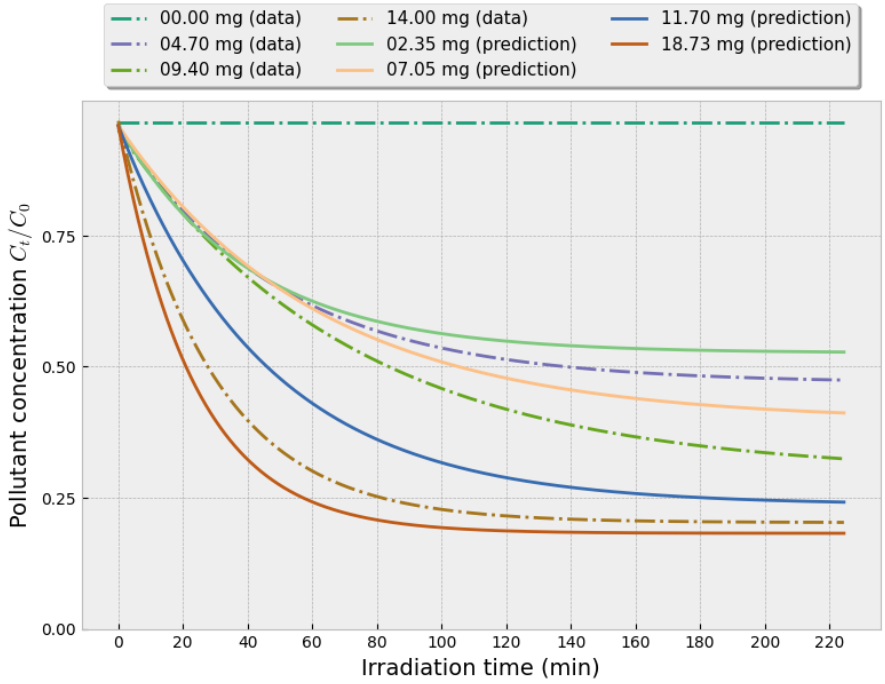


(a) Sample of predictions

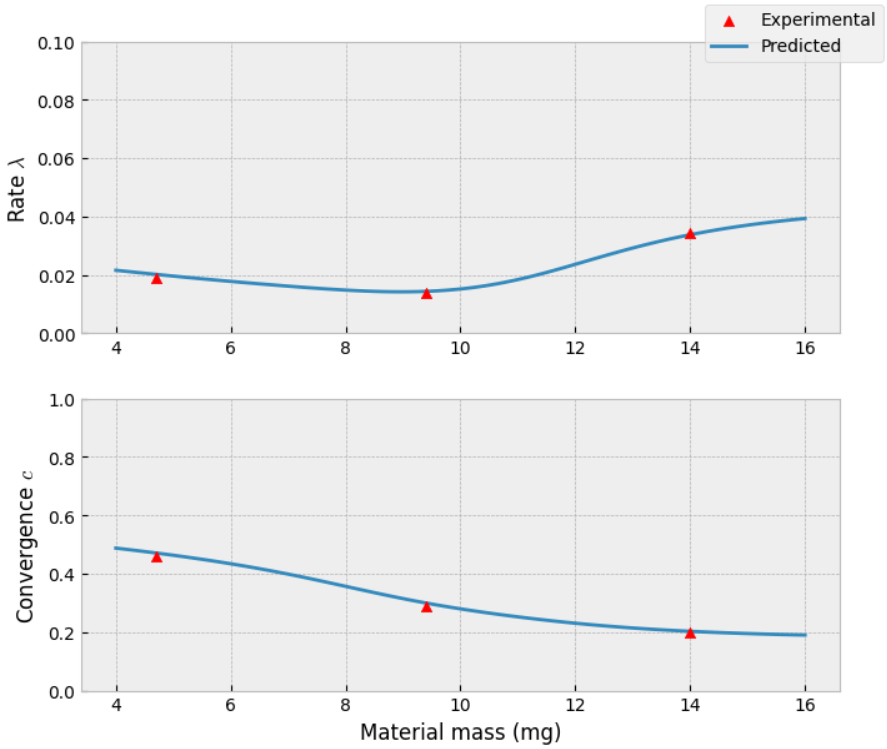


(b) Rate and convergence

FIGURE 3.7: Base model trained on the change of initial pollutant concentration (TiO_2/curcumin material mass = 14 mg, light intensity = 3.76 w/cm²)



(a) Sample of predictions



(b) Rate and convergence

FIGURE 3.8: Base model trained on the change of material mass (Light intensity = 3.76 w/cm², initial MB concentration = 10 ppm)

3.10.3 Degree of extrapolation

For each base model, the upper and lower bounds as well as the number of predictions were selected appropriately, as shown in the following table 3.1:

Base Model	$\mathbf{s}_{\text{min}}$	$\mathbf{s}_{\text{max}}$	n_preds
Light intensity	0	5	15
Initial pollutant concentration	0	20	15
Material mass	0	16	15

TABLE 3.1: Range of extrapolation and number of predictions by base model

3.10.4 Hyperparameter tuning

The base and meta models underwent comprehensive tuning with a grid-search algorithm, the scoring was guided by MSE loss. Various activation functions were evaluated independently for the hidden and output layer: sigmoid, tanh, ReLU. Different dense layer topologies were selected: (128, 64), (64, 32), (32, 16), and (16, 8). For the meta model, the number of units h of the weight matrices included the values: 7, 16, 32, and 64. The resulting augmented dataset, $\mathcal{A}'$, was shuffled and used for both the training and validation of the meta model. The best hyperparameter configurations for the base and meta models are shown in table 3.2.

Parameter	**Hidden act.**	**Output act.**	**Topology**
Base model	Sigmoid	Sigmoid	(64, 32, 16, 2)
Meta model	ReLU	Sigmoid	(h = 7) | (32, 16, 2)

TABLE 3.2: Optimal hyperparameters across the models of the ensemble

With these configurations, the base models achieved good fits to the experimental data, accurately capturing the degradation rates and convergence (see figure 3.6, 3.7, 3.8). The loss scores for the base models on their respective subsets, as well as for the meta model, are shown in table 3.3.

	MSE	**MAE**	**RMSE**
Light intensity model	3.03e-6	0.000843	0.00174
Initial pollutant concentration model	1.66e-5	0.00294	0.00407
Material mass model	4.99e-6	0.00114	0.00223
Meta model	6.00e-5	0.0057	0.0079

TABLE 3.3: Loss scores across the models of the ensemble

3.10.5 Validation

The meta model was validated on the augmented dataset $\mathcal{A}'$ with a split of 80% for training and 20% for validation using a 5-fold cross-validation. The performance metric used for this validation was MSE and the resulting score was 0.0055.

3.10.6 Predictions

A grid-search was employed through different configurations of experimental parameters in order to identify the optimal combination. The range was dictated by $s^j_{\min}$ and $s^j_{\max}$. The model predicted optimal parameter ranges that resulted in efficiencies of 86% to 90%.

Experimental parameter	**Optimal range**
Material mass (mg)	[20, 22]
Light intensity (w/cm²)	[4.4, 5.7]
Initial MB concentration (ppm)	[11, 15]

TABLE 3.4: Optimal range by experimental parameter

3.11 Conclusion

The presented DASAN model shows a promising approach for predicting the photocatalytic degradation rate and efficiency from experimental parameters. Despite the lack of data, it demonstrated high accuracy with a mean squared error of 0.0055 through 5-fold cross-validation. However, the approach relies heavily on the specific experimental methodology used to obtain the data. Future research could expand the approach to larger and more appropriate datasets, explore different photocatalyst materials, and incorporate additional experimental factors. This study underscores the potential of data-augmenting ensembles in guiding experimental design for photocatalysis, particularly in data-limited scenarios.

General conclusion

This thesis aimed to explore the effectiveness of data-augmenting self-attention ensemble models in the context of photocatalytic degradation processes

The study was organized into three core chapters. The first chapter established a comprehensive theoretical background on modeling and simulation, with a particular focus on complex systems and their classifications. The second chapter delved into the fundamentals and techniques of machine learning and deep learning and provided an overview of artificial neural networks and attention mechanisms. The third chapter presented the implementation of the ensemble and detailed the experimental setup, dataset, evaluation metrics, and results.

Although addressing challenges such as data limitations and dependence on experimental methodology remains a hurdle that requires further research for model optimization, the study concluded that the self-attention ensemble model shows promise in enhancing photocatalytic degradation predictions.

Bibliography

[1] Sughosh Madhav et al. "Water Pollutants: Sources and Impact on the Environment and Human Health". In: *Sensors in Water Pollutants Monitoring: Role of Material*. Singapore: Springer Singapore, 2020, pp. 43–62. ISBN: 978-981-15-0671-0. DOI: 10.1007/978-981-15-0671-0_4. URL: https://doi.org/10.1007/978-981-15-0671-0_4.

[2] Robert I. McDonald et al. "Water on an urban planet: Urbanization and the reach of urban water infrastructure". In: *Global Environmental Change* 27 (2014), pp. 96–105. ISSN: 0959-3780. DOI: https://doi.org/10.1016/j.gloenvcha.2014.04.022. URL: https://www.sciencedirect.com/science/article/pii/S0959378014000880.

[3] Anil K Dwivedi. "Researches in water pollution: A review". In: *International Research Journal of Natural and Applied Sciences* 4.1 (2017), pp. 118–142.

[4] Shilpi Das, Susmita Mishra, and Himadri Sahu. "A review of activated carbon to counteract the effect of iron toxicity on the environment". In: *Environmental Chemistry and Ecotoxicology* 5 (2023), pp. 86–97. ISSN: 2590-1826. DOI: https://doi.org/10.1016/j.enceco.2023.02.002. URL: https://www.sciencedirect.com/science/article/pii/S259018262300005X.

[5] Guy Ryder et al. *The United Nations world water development report, 2017: Wastewater: the untapped resource*. 2017.

[6] Natural Resources Defense Council. *Encourage textile manufacturers to reduce pollution*. Mar. 7, 2024. URL: https://www.nrdc.org/resources/encourage-textile-manufacturers-reduce-pollution.

[7] Bruno Lellis et al. "Effects of textile dyes on health and the environment and bioremediation potential of living organisms". In: *Biotechnology Research and Innovation* 3.2 (2019), pp. 275–290. ISSN: 2452-0721. DOI: https://doi.org/10.1016/j.biori.2019.09.001. URL: https://www.sciencedirect.com/science/article/pii/S2452072119300413.

[8] Michael C. Petriello et al. "Modulation of persistent organic pollutant toxicity through nutritional intervention: Emerging opportunities in biomedicine and environmental remediation". In: *Science of The Total Environment* 491-492 (2014). Halogenated Persistent Organic Pollutants (Dioxin2013, Daegu/Korea), pp. 11–16. ISSN: 0048-9697. DOI: https://doi.org/10.1016/j.scitotenv.2014.01.109. URL: https://www.sciencedirect.com/science/article/pii/S0048969714001387.

[9] Duk-Hee Lee et al. "Chlorinated persistent organic pollutants, obesity, and type 2 diabetes". en. In: *Endocr Rev* 35.4 (Jan. 2014), pp. 557–601.

[10] Qing Qing Li et al. "Persistent Organic Pollutants and Adverse Health Effects in Humans". In: *Journal of Toxicology and Environmental Health, Part A* 69.21 (2006). PMID: 16982537, pp. 1987–2005. DOI: 10.1080/15287390600751447. eprint: https://doi.org/10.1080/15287390600751447. URL: https://doi.org/10.1080/15287390600751447.

[11] F. Wania, J. Axelman, and D. Broman. "A review of processes involved in the exchange of persistent organic pollutants across the air–sea interface". In: *Environmental Pollution* 102.1 (1998), pp. 3–23. ISSN: 0269-7491. DOI: https://doi.org/10.1016/S0269-7491(98)00072-4. URL: https://www.sciencedirect.com/science/article/pii/S0269749198000724.

[12] Jolly Jacob and Jacob Cherian. "Review of environmental and human exposure to persistent organic pollutants". In: *Asian Social Science* 9.11 (2013), p. 107.

[13] World Health Organization et al. *Burden of disease attributable to unsafe drinking-water, sanitation and hygiene*. World Health Organization, 2023.

[14] Sanjay K. Sharma, Rashmi Sanghi, and Ackmez Mudhoo. "Green Practices to Save Our Precious "Water Resource"". In: *Advances in Water Treatment and Pollution Prevention*. Dordrecht: Springer Netherlands, 2012, pp. 1–36. ISBN: 978-94-007-4204-8. DOI: 10.1007/978-94-007-4204-8_1. URL: https://doi.org/10.1007/978-94-007-4204-8_1.

[15] M. Abdennouri et al. "Photocatalytic degradation of pesticides by titanium dioxide and titanium pillared purified clays". In: *Arabian Journal of Chemistry* 9 (2016), S313–S318. ISSN: 1878-5352. DOI: https://doi.org/10.1016/j.arabjc.2011.04.005. URL: https://www.sciencedirect.com/science/article/pii/S1878535211001109.

[16] Muhammad Saeed et al. "Photocatalysis: an effective tool for photodegradation of dyes—a review". In: *Environmental Science and Pollution Research* 29.1 (2022), pp. 293–311. ISSN: 1614-7499. DOI: 10.1007/s11356-021-16389-7. URL: https://doi.org/10.1007/s11356-021-16389-7.

[17] Weng Shin Koe et al. "An overview of photocatalytic degradation: photocatalysts, mechanisms, and development of photocatalytic membrane". In: *Environmental Science and Pollution Research* 27.3 (2020), pp. 2522–2565. ISSN: 1614-7499. DOI: 10.1007/s11356-019-07193-5. URL: https://doi.org/10.1007/s11356-019-07193-5.

[18] S. Chandhini Priya et al. "A critical review on efficient photocatalytic degradation of organic compounds using copper-based nanoparticles". In: *Materials Today: Proceedings* 80 (2023). SI:5 NANO 2021, pp. 3075–3081. ISSN: 2214-7853. DOI: https://doi.org/10.1016/j.matpr.2021.07.169. URL: https://www.sciencedirect.com/science/article/pii/S2214785321050288.

[19] Xiangchao Meng, Nan Yun, and Zisheng Zhang. "Recent advances in computational photocatalysis: A review". In: *The Canadian Journal of Chemical Engineering* 97.7 (2019), pp. 1982–1998. DOI: https://doi.org/10.1002/cjce.23477. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cjce.23477. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/cjce.23477.

[20] Tânia F. G. G. Cova and Alberto A. C. C. Pais. *Deep Learning for Deep Chemistry: Optimizing the Prediction of Chemical Patterns*. Nov. 2019. DOI: 10.3389/fchem.2019.00809. URL: http://dx.doi.org/10.3389/fchem.2019.00809.

[21] Adam C. Mater and Michelle L. Coote. "Deep Learning in Chemistry". In: *Journal of Chemical Information and Modeling* 59.6 (2019), pp. 2545–2559. DOI: 10.1021/acs.jcim.9b00266. eprint: https://doi.org/10.1021/acs.jcim.9b00266. URL: https://doi.org/10.1021/acs.jcim.9b00266.

[22] Rafael Gómez-Bombarelli et al. "Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules". In: *ACS Central Science* 4.2 (2018). PMID: 29532027, pp. 268–276. DOI: 10.1021/acscentsci.7b00572. eprint: https://doi.org/10.1021/acscentsci.7b00572. URL: https://doi.org/10.1021/acscentsci.7b00572.

[23] Matthew A Kayala and Pierre Baldi. "ReactionPredictor: prediction of complex chemical reactions at the mechanistic level using machine learning". en. In: *J Chem Inf Model* 52.10 (Oct. 2012), pp. 2526–2540.

[24] Alexandru Korotcov et al. "Comparison of Deep Learning With Multiple Machine Learning Methods and Metrics Using Diverse Drug Discovery Data Sets". en. In: *Mol Pharm* 14.12 (Nov. 2017), pp. 4462–4475.

[25] Gianni Brauwers and Flavius Frasincar. "A General Survey on Attention Mechanisms in Deep Learning". In: *IEEE Transactions on Knowledge and Data Engineering* 35.4 (2023), pp. 3279–3298. DOI: 10.1109/TKDE.2021.3126456.

[26] Asmaa Bouazza et al. "Use of TiO2/curcumin nanocomposite material deposited on a cellulosic film for methylene blue photocatalytic degradation under UV light". In: *Reaction Kinetics, Mechanisms and Catalysis* 136.3 (2023), pp. 1625–1641. ISSN: 1878-5204. DOI: 10.1007/s11144-023-02429-5. URL: https://doi.org/10.1007/s11144-023-02429-5.

[27] Hiroki Sayama. *Introduction to the modeling and analysis of complex systems*. Open SUNY Textbooks, 2015.

[28] James Ladyman, James Lambert, and Karoline Wiesner. "What is a complex system?" In: *European Journal for Philosophy of Science* 3.1 (Jan. 2013), pp. 33–67.

[29] David Adam. "Special report: The simulations driving the world's response to COVID-19." In: *Nature* 580.7802 (2020), pp. 316–319.

[30] Michael J Grimble et al. "Introduction to nonlinear systems modelling and control". In: *Nonlinear Industrial Control Systems: Optimal Polynomial Systems and State-Space Approach* (2020), pp. 3–63.

[31] Bernard P Zeigler, Herbert Praehofer, and Tag Gon Kim. *Theory of modeling and simulation*. Academic press, 2000.

[32] Satvir Singh, Arun Khosla, and JS Saini. "Nature-Inspired Toolbox to Design and Optimize Systems". In: *Machine Learning Algorithms for Problem Solving in Computational Applications: Intelligent Techniques*. IGI Global, 2012, pp. 273–291.

[33] Ahmet Erdemir et al. "Credible practice of modeling and simulation in healthcare: ten rules from a multidisciplinary perspective". In: *Journal of translational medicine* 18.1 (2020), p. 369.

[34] Stefan Riedmaier et al. "Unified framework and survey for model verification, validation and uncertainty quantification". In: *Archives of Computational Methods in Engineering* 28 (2021), pp. 2655–2688.

[35] David C Wynn and Claudia M Eckert. "Perspectives on iteration in design and development". In: *Research in Engineering Design* 28 (2017), pp. 153–184.

[36] National Institute of Biomedical Imaging and Bioengineering. *Computational Modeling | National Institute of Biomedical Imaging and Bioengineering*. Nih.gov, 2009. URL: https://www.nibib.nih.gov/science-education/science-topics/computational-modeling (visited on 05/31/2024).

[37] George J Klir. *Architecture of systems problem solving*. Springer Science & Business Media, 2013.

[38] Bernard P. Zeigler. "Extending the Hierarchy of System Specifications and Morphisms with SES Abstraction". In: *Information* 14.1 (2023). ISSN: 2078-2489. DOI: 10.3390/info14010022. URL: https://www.mdpi.com/2078-2489/14/1/22.

[39] Sangeeth S Ponnusamy, Vincent Albert, and Patrice Thebault. "Consistent behavioral abstractions of experimental frame". In: *AIAA Modeling and Simulation Technologies Conference*. 2016, p. 1923.

[40] Deniz Cetinkaya, Alexander Verbraeck, and Mamadou Seck. "Applying a model driven approach to component based modeling and simulation". In: Dec. 2010, pp. 546–553. DOI: 10.1109/WSC.2010.5679131.

[41] Catherine L Lawson et al. "Cryo-EM model validation recommendations based on outcomes of the 2019 EMDataResource challenge". In: *Nature methods* 18.2 (2021), pp. 156–164.

[42] N. Nilsson. "Principles of Artificial Intelligence". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-3 (1980), pp. 112–112. DOI: 10.1007/978-3-662-09438-9.

[43] Tom M Mitchell. *Machine learning*. Vol. 1. 9. McGraw-hill New York, 1997.

[44] Qian Zhang, Jie Lu, and Yaochu Jin. "Artificial intelligence in recommender systems". In: *Complex & Intelligent Systems* 7 (2020), pp. 439–457. DOI: 10.1007/s40747-020-00212-w.

[45] H. Salehi and R. Burgueño. "Emerging artificial intelligence methods in structural engineering". In: *Engineering Structures* (2018). DOI: 10.1016/J.ENGSTRUCT.2018.05.084.

[46] Fei Wang and A. Preininger. "AI in Health: State of the Art, Challenges, and Future Directions". In: *Yearbook of Medical Informatics* 28 (2019), pp. 16 –26. DOI: 10.1055/s-0039-1677908.

[47] Longbing Cao. "AI in Finance: Challenges, Techniques, and Opportunities". In: *ACM Computing Surveys (CSUR)* 55 (2021), pp. 1 –38. DOI: 10.1145/3502289.

[48] Daniele Ravì et al. "Deep learning for health informatics". In: *IEEE journal of biomedical and health informatics* 21.1 (2016), pp. 4–21.

[49] Shimin Su et al. "Predicting the feasibility of copper (i)-catalyzed alkyne–azide cycloaddition reactions using a recurrent neural network with a self-attention mechanism". In: *Journal of Chemical Information and Modeling* 60.3 (2020), pp. 1165–1174.

[50] M. Kubát. "An Introduction to Machine Learning". In: (2017), pp. 1–348. DOI: 10.1007/978-3-319-63913-0.

[51] P. E. Miller. "Predictive Abilities of Machine Learning Techniques May Be Limited by Dataset Characteristics: Insights From the UNOS Database." In: *Journal of cardiac failure* (2019).

[52] P. Jonathon Phillips et al. *Four Principles of Explainable Artificial Intelligence*. en. 2021. DOI: https://doi.org/10.6028/NIST.IR.8312. URL: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=933399.

[53] Ninareh Mehrabi et al. "A Survey on Bias and Fairness in Machine Learning". In: *ACM Comput. Surv.* 54.6 (2021). ISSN: 0360-0300. DOI: 10.1145/3457607. URL: https://doi.org/10.1145/3457607.

[54] Stuart J Russell and Peter Norvig. *Artificial intelligence: a modern approach*. Pearson, 2016.

[55] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[56] Warren S. McCulloch and Walter Pitts. "A logical calculus of the ideas immanent in nervous activity". In: *The bulletin of mathematical biophysics* 5.4 (1943), pp. 115–133. ISSN: 1522-9602. DOI: 10.1007/BF02478259. URL: https://doi.org/10.1007/BF02478259.

[57] J.G. Betts. *Anatomy and Physiology*. Open Textbook Library. OpenStax College, Rice University, 2013. ISBN: 9781938168130. URL: https://books.google.dz/books?id=dvVgngEACAAJ.

[58] Vinod Nair and Geoffrey E Hinton. "Rectified linear units improve restricted boltzmann machines". In: *ICML 2010*. 2010, pp. 807–814.

[59] B.L. Kalman and S.C. Kwasny. "Why tanh: choosing a sigmoidal function". In: *[Proceedings 1992] IJCNN International Joint Conference on Neural Networks*. Vol. 4. 1992, 578–581 vol.4. DOI: 10.1109/IJCNN.1992.227257.

[60] Dan Hendrycks and Kevin Gimpel. *Gaussian Error Linear Units (GELUs)*. 2023. arXiv: 1606.08415 [cs.LG].

[61] Prajit Ramachandran, Barret Zoph, and Quoc V. Le. *Searching for Activation Functions*. 2017. arXiv: 1710.05941 [cs.NE].

[62] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: (2017). arXiv: 1412.6980 [cs.LG].

[63] Manzil Zaheer et al. "Adaptive Methods for Nonconvex Optimization". In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio et al. Vol. 31. Curran Associates, Inc., 2018. URL: https://proceedings.neurips.cc/paper_files/paper/2018/file/90365351ccc7437a1309dc64e4db3Paper.pdf.

[64] T. Tieleman. "Lecture 6.5-rmsprop: Divide the Gradient by a Running Average of Its Recent Magnitude". In: *COURSERA: Neural Networks for Machine Learning* 4.2 (2012), p. 26. URL: https://cir.nii.ac.jp/crid/1370017282431050757.

[65] B.T. Polyak. "Some methods of speeding up the convergence of iteration methods". In: *USSR Computational Mathematics and Mathematical Physics* 4.5 (1964), pp. 1–17. ISSN: 0041-5553. DOI: https://doi.org/10.1016/0041-5553(64)90137-5. URL: https://www.sciencedirect.com/science/article/pii/0041555364901375.

[66] Zhaoyang Niu, Guoqiang Zhong, and Hui Yu. "A review on the attention mechanism of deep learning". In: *Neurocomputing* 452 (2021), pp. 48–62. ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2021.03.091. URL: https://www.sciencedirect.com/science/article/pii/S092523122100477X.

[67] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. *Neural Machine Translation by Jointly Learning to Align and Translate*. 2016. arXiv: 1409.0473 [cs.CL].

[68] Aston Zhang et al. *Dive into Deep Learning*. https://D2L.ai. Cambridge University Press, 2023.

[69] Ashish Vaswani et al. *Attention Is All You Need*. 2023. arXiv: 1706.03762 [cs.CL].

[70] Aston Zhang et al. *Dive into Deep Learning*. https://D2L.ai. Cambridge University Press, 2023.

[71] Zhi-Hua Zhou. "Ensemble Learning". In: *Encyclopedia of Biometrics*. Ed. by Stan Z. Li and Anil Jain. Boston, MA: Springer US, 2009, pp. 270–273. ISBN: 978-0-387-73003-5. DOI: 10.1007/978-0-387-73003-5_293. URL: https://doi.org/10.1007/978-0-387-73003-5_293.

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

[73] Yoav Freund and Robert E Schapire. "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting". In: *Journal of Computer and System Sciences* 55.1 (1997), pp. 119–139. ISSN: 0022-0000. DOI: https://doi.org/10.1006/jcss.1997.1504. URL: https://www.sciencedirect.com/science/article/pii/S002200009791504X.

[74] David H. Wolpert. "Stacked generalization". In: *Neural Networks* 5.2 (1992), pp. 241–259. ISSN: 0893-6080. DOI: https://doi.org/10.1016/S0893-6080(05)80023-1. URL: https://www.sciencedirect.com/science/article/pii/S0893608005800231.

[75] *What is Stacking in Machine Learning?- Scaler Topics*. URL: https://www.scaler.com/topics/machine-learning/stacking-in-machine-learning.

[76] N.R. Khalid et al. "Carbonaceous-TiO2 nanomaterials for photocatalytic degradation of pollutants: A review". In: *Ceramics International* 43.17 (2017), pp. 14552–14571. ISSN: 0272-8842. DOI: https://doi.org/10.1016/j.ceramint.2017.08.143. URL: https://www.sciencedirect.com/science/article/pii/S0272884217318400.

[77] Mei Han et al. "Recent progress on the photocatalysis of carbon dots: Classification, mechanism and applications". In: *Nano Today* 19 (2018), pp. 201–218. ISSN: 1748-0132. DOI: https://doi.org/10.1016/j.nantod.2018.02.008. URL: https://www.sciencedirect.com/science/article/pii/S1748013217305777.

[78] Dhananjay S Bhatkhande, Vishwas G Pangarkar, and Anthony A C M Beenackers. "Photocatalytic degradation for environmental applications – a review". In: *Journal of Chemical Technology & Biotechnology* 77.1 (2002), pp. 102–116. DOI: https://doi.org/10.1002/jctb.532. eprint: https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/pdf/10.1002/jctb.532. URL: https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/jctb.532.

[79] Ankit Kumar and G Pandey. "A review on the factors affecting the photocatalytic degradation of hazardous materials". In: *Mater. Sci. Eng. Int. J* 1.3 (2017), pp. 1–10.

[80] Shrabana Sarkar et al. "Green polymeric nanomaterials for the photocatalytic degradation of dyes: a review". In: *Environmental Chemistry Letters* 18.5 (2020), pp. 1569–1580.

[81] Safa Al-Yahyaey et al. "Multi-channel flow reactor design for the photocatalytic degradation of harmful dye molecules". In: *Journal of Nanoparticle Research* 26.4 (2024), p. 72. ISSN: 1572-896X. DOI: 10.1007/s11051-024-05981-w. URL: https://doi.org/10.1007/s11051-024-05981-w.

[82] Abbas Al-Nayili and Wissam A. Alhaidry. "Batch to continuous photocatalytic degradation of phenol using nitrogen-rich g-C3N4 nanocomposites". In: *Research on Chemical Intermediates* 49.10 (2023), pp. 4239–4255. ISSN: 1568-5675. DOI: 10.1007/s11164-023-05099-z. URL: https://doi.org/10.1007/s11164-023-05099-z.

[83] Susmita Das et al. "Artificial neural network modeling of photocatalytic degradation of pollutants: a review of photocatalyst, optimum parameters and model topology". In: *Catalysis Reviews* 0.0 (2024), pp. 1–35. DOI: 10.1080/01614940.2024.2338131. eprint: https://doi.org/10.1080/01614940.2024.2338131. URL: https://doi.org/10.1080/01614940.2024.2338131.

[84] Zhuoying Jiang et al. "A generalized predictive model for TiO2–Catalyzed photo-degradation rate constants of water contaminants through artificial neural network". In: *Environmental Research* 187 (2020), p. 109697. ISSN: 0013-9351. DOI: https://doi.org/10.1016/j.envres.2020.109697. URL: https://www.sciencedirect.com/science/article/pii/S0013935120305909.

[85] Garg Alok et al. "Optimization methodology based on neural networks and box-behnken design applied to photocatalysis of acid red 114 dye". In: *Environmental Engineering Research* 25.5 (2020), pp. 753–762. DOI: 10.4491/eer.2019.246. eprint: http://www.eeer.org/journal/view.php?number=1097. URL: http://www.eeer.org/journal/view.php?number=1097.

[86] Chandrika K.C et al. "Applications of artificial neural network and Box-Behnken Design for modelling malachite green dye degradation from textile effluents using TiO2 photocatalyst". In: *Environmental Engineering Research* 27 (Jan. 2021). DOI: 10.4491/eer.2020.553.

[87] B. Boutra, A. Sebti, and M. Trari. "Response surface methodology and artificial neural network for optimization and modeling the photodegradation of organic pollutants in water". In: *International Journal of Environmental Science and Technology* 19.11 (2022), pp. 11263–11278. ISSN: 1735-2630. DOI: 10.1007/s13762-021-03875-1. URL: https://doi.org/10.1007/s13762-021-03875-1.