Democratic and Popular Republic of Algeria

Ministry of Higher Education and Scientific Research

Hassiba Ben Bouali University – Chlef

Faculty of Technology

Department of Process Engineering

Final Year Project

Towards a Master's Degree in Process Engineering

Specialty: Pharmaceutical Engineering

Theme

Study and modeling by artificial intelligence of occupational exposure limits to certain pharmaceutical products using chemoinformatics

Presented by:

MORSLI Yamina

Directed by:

Dr. HENTABLI Mohamed

Before the jury composed of:

- President**: Dr ALLICHE**
- Examiner: **Dr AYADI**
- Supervisor: **Dr HENTABLI**

Schoolar year: **2024/2025**


بسم الله الرحمن الرحيم

Acknowledgements

*First and foremost, I would like to express my deepest gratitude to **ALLAH** for granting me the strength, patience, and guidance to complete this thesis. Without His divine grace and infinite blessings, this journey would not have been possible.*

*I would also like to extend my sincere thanks to all those who, in one way or another, contributed to the successful completion of this Master's thesis in **Pharmaceutical Engineering**.*

*I am especially grateful to **Mr. Mohamed Hentabli**, my thesis supervisor, for his availability, kindness, insightful guidance, and continuous support throughout this work. His expertise in the pharmaceutical field has been an invaluable source of knowledge and has greatly contributed to both my academic and personal development.*

*My heartfelt appreciation also goes to **Mr. Alliche** and **Miss Ayadi** for honoring me by serving as members of my thesis jury. I am truly thankful for your time, constructive feedback, and thoughtful evaluation of my work.*

*I would like to acknowledge all the professors and lecturers of the Master's program in Pharmaceutical Engineering at the **Faculty of Technology, Hassiba Ben Bouali University**. Your dedication to teaching, academic rigor, and the knowledge you have shared have played a significant role in shaping my academic path.*

To my classmates, colleagues, and friends — thank you for your support, collaboration, and the many enriching exchanges we have shared throughout this journey.

To all of you — and above all, to God — thank you from the bottom of my heart.

Yamina

Dedications

This thesis is whole heartedly dedicated to my beloved family — the pillars of my strength, motivation, and success.

To my dear parents, ***Abdelkader and Lalia****,*

Your unconditional love, sacrifices, and constant encouragement have been my guiding light throughout this journey. You have taught me the true meaning of perseverance and dedication. Your unwavering belief in me and your endless support have made this achievement possible. I am forever grateful for your presence in my life.

To my brothers and sisters,

To ***Ibrahim****, my brother, thank you for your steadfast support and encouragement. Your confidence in me has always been a source of motivation.*

To ***Hadjer*** *and* ***Saba****, my sisters, your kindness, understanding, and words of comfort have lifted my spirits in moments of doubt. I am deeply thankful to have you by my side.*

To my entire family, your love and faith in me have been a foundation on which I have built my dreams.

This work is not only my accomplishment but also a tribute to the love, values, and strength you have all given me.

I dedicate this thesis to you, with all my heart and deepest gratitude.

Abstract

Artificial intelligence is increasingly being utilized to all field and specially to enhance occupational health within the pharmaceutical industry, where the rising potency of compounds contributes to elevated exposure risks. Due to limited toxicity data for many emerging drug candidates, traditional methods often fall short in accurately estimating safe handling thresholds. In this study, an AI-driven approach has been developed to predict Occupational Exposure Bands (OEBs) based on molecular structures. This method combines cheminformatics descriptors and molecular fingerprints with deep learning techniques to extract significant features, which are then classified using various machine learning algorithms. The resulting models exhibit strong predictive performance, although challenges remain in accurately identifying less-represented high-risk categories. Additionally, a practical software tool was created to facilitate real-time OEB predictions and molecular visualization, providing an accessible interface for researchers and safety professionals. Overall, this approach offers an innovative solution for early hazard assessment and highlights the potential of AI to improve workplace safety in pharmaceutical development.

Key words: Artificial Intelligence; Pharmaceutical Industry; Occupational Exposure Bands (OEBs); Molecular Structures; Cheminformatics; Deep Learning; Machine Learning; Toxicity Prediction.

Résumé

L'intelligence artificielle est de plus en plus utilisée dans tous les domaines, et plus particulièrement pour améliorer la santé au travail dans l'industrie pharmaceutique, où la puissance croissante des composés contribue à des risques d'exposition élevés. En raison du manque de données de toxicité pour de nombreux candidats médicaments émergents, les méthodes traditionnelles ne permettent souvent pas d'estimer avec précision les seuils de manipulation sécuritaire. Dans cette étude, une approche basée sur l'IA a été développée pour prédire les bandes d'exposition professionnelle (BEP) à partir de structures moléculaires. Cette méthode combine des descripteurs chimioinformatiques et des empreintes moléculaires avec des techniques d'apprentissage profond pour extraire des caractéristiques significatives, qui sont ensuite classées à l'aide de divers algorithmes d'apprentissage automatique. Les modèles obtenus affichent d'excellentes performances prédictives, bien que l'identification précise des catégories à haut risque moins représentées pose encore problème. De plus, un outil logiciel pratique a été créé pour faciliter les prédictions des BEP en temps réel et la visualisation moléculaire, offrant une interface accessible aux chercheurs et aux professionnels de la sécurité. Globalement, cette approche offre une solution innovante pour l'évaluation précoce des dangers et souligne le potentiel de l'IA pour améliorer la sécurité au travail dans le développement pharmaceutique.

Mots clés : Intelligence artificielle ; Industrie pharmaceutique ; Bandes d'exposition professionnelle (BEP) ; Structures moléculaires ; chimioinformatique ; Apprentissage profond ; apprentissage automatique ; prédiction de toxicité.

الملخص

يُستخدم الذكاء الاصطناعي بشكل متزايد في مختلف المجالات، وخاصة لتعزيز الصحة المهنية في صناعة الأدوية، حيث يؤدي ارتفاع فعالية المركبات إلى زيادة مخاطر التعرض. ونظرًا لندرة البيانات السمية المتاحة للعديد من المرشحين الدوائيين الجدد، غالبًا ما تكون الطرق التقليدية غير كافية لتقدير مستويات التعرض الآمنة بدقة .في هذا العمل، تم تطوير نهج يعتمد على الذكاء الاصطناعي لتوقع نطاقات التعرض المهني(OEBs) بناءً على البنية الجزيئية للمركبات. يجمع هذا النهج بين أوصاف المعلومات الكيميائية وبصمات الجزيئات مع تقنيات التعلم العميق لاستخلاص الميزات المهمة، والتي يتم تصنيفها باستخدام خوارزميات تعلم آلي متنوعة. أظهرت النماذج أداءً تنبؤيًا قويًا، مع وجود تحديات في التعرف بدقة على الفئات عالية الخطورة التي تمثل بنسبة أقل .بالإضافة إلى ذلك، تم إنشاء أداة برمجية عملية لتسهيل التنبؤات الفورية لنطاقات التعرض وعرض البنية الجزيئية، مما يوفر واجهة سهلة الاستخدام للباحثين والمتخصصين في مجال السلامة المهنية. بشكل عام، يقدم هذا النهج حلاً مبتكرًا لتقييم المخاطر المبكرة ويبرز قدرة الذكاء الاصطناعي على تحسين السلامة في بيئات تطوير الأدوية .

الكلمات المفتاحية: الذكاء الاصطناعي، صناعة الأدوية، نطاقات التعرض المهني، البنية الجزيئية، المعلومات الكيميائية، التعلم العميق، خوارزميات التعلم الآلي، التنبؤ بالسمية.

Table of contents

List of figures

List of tables

List of abbreviations:

GHS: globaly harmonized system

ODs: occupational Diseases

OELs: occupational exposure limits

ACGIH:

TWA: time weighted average

STEL: short term exposure limits

SDS: safety data sheet

AI : artificiel intelligence

ML: machine learning

DL: deep learning

ANNs: artificial neural network

CNNs : conventionnel neural network

RNN: recurrent neural network

FNN: feed Forword neural network

MLPs: multilayer perceptron's

BP: back propagation

MDs: molecular descritors

QSAR: quantity structure activity relationship

OEB: occupational exposure bands

GUI: graphical user interface

INTRODUCTION

Introduction:

In recent years, the pharmaceutical industry has witnessed a paradigm shift in drug development, transitioning from a primary focus on small-molecule drugs to the exploration of advanced therapeutic modalities such as peptides, nucleic acids (DNA/RNA), and biologics. These emerging therapies, designed for greater potency and precision, often require lower doses and less frequent administration. However, this increased potency also brings heightened risks of occupational exposure, even at minimal concentrations. Such exposure poses significant health hazards not only to workers involved in drug synthesis and manufacturing but also to healthcare providers, patients, and others sharing clinical or domestic environments.

Despite the establishment of regulatory frameworks like Safety Data Sheets (SDS), the Globally Harmonized System (GHS), and the National Institute for Occupational Safety and Health (NIOSH) hazardous drug lists, there remains a pressing need for more proactive and predictive tools to manage the risks posed by potent pharmaceutical compounds—especially those still in early development. Many of these compounds lack comprehensive nonclinical or clinical toxicity data. In such cases, occupational exposure banding (OEB) systems are employed to estimate acceptable exposure ranges and necessary safety measures. These banding decisions often rely on in silico models, read-across strategies, in vitro assays, and, where available, in vivo data.

Artificial Intelligence (AI), with its unparalleled ability to process and analyze large, complex datasets, offers a transformative solution to the challenge of predicting occupational exposure risks. In the pharmaceutical domain, AI has already demonstrated its value in accelerating drug discovery, optimizing molecular design, and reducing reliance on animal testing. Now, its potential is being extended to the field of occupational health.

By leveraging machine learning and chemoinformatics, AI can be used to predict Occupational Exposure Limits (OELs) for chemical substances, particularly those lacking sufficient empirical data. AI-driven models can rapidly identify molecular patterns associated with toxicity, correlate chemical structure with potential biological effects, and provide early risk assessments. This predictive capacity not only enhances worker safety but also supports regulatory compliance, risk management planning, and cost-effective safety evaluation.

Moreover, AI technologies can be integrated into real-time environmental monitoring systems, using smart sensors to continuously track air quality in manufacturing facilities. These systems can trigger instant alerts when exposure thresholds are breached, enabling swift intervention

and minimizing health hazards. AI also aids in the development of personalized protective measures, dynamic risk mitigation strategies, and responsive emergency protocols—all grounded in both real-time and historical data analysis.

This research seeks to contribute to occupational health and safety by developing AI-based models capable of predicting OELs for pharmaceutical compounds that lack well-established exposure data. By combining artificial intelligence with chemoinformatics, we aim to create a robust framework that supports early hazard identification, informed decision-making, and sustainable workplace safety practices.

To achieve these objectives, the study is structured into two main parts:

Theoretical Section

Chapter 1 outlines fundamental concepts of occupational exposure, defines key diseases associated with pharmaceutical exposure, and explains the significance and regulatory background of Occupational Exposure Limits.

Chapter 2 presents the AI techniques, chemoinformatics tools, and modeling algorithms used in this study, including their relevance and implementation.

Experimental Section

Chapter 3 showcases the practical application of these methods, detailing the results, analysis, and discussion based on the developed predictive models.

Through this work, we aim to bridge the gap between cutting-edge technology and occupational health, ultimately promoting a safer and more responsible pharmaceutical manufacturing environment.

Theoretical part

CHAPTER I

OCCUPATIONAL DISEASE AND OCCUPATIONAL EXPOSURE LIMITS

1. Occupational diseases:

1.1. Definition:

Several definitions of the term "occupational disease" exist. However, for the purpose of the Protocol of 2002 to the Occupational Safety and Health Convention of International Labour Organisation (ILO), the term 'occupational disease' covers any disease contracted as a result of an exposure to risk factors arising from work activity". The ILO Employment Injury Benefits Recommendation defines occupational diseases more precisely in the following terms: "Each Member should, under prescribed conditions, regard diseases known to arise out of the exposure to substances and dangerous conditions in processes, trades or occupations as occupational diseases [1].

Figure I.1: Occupational hazards

1.2. Route to expose by chemicals agents:

Chemicals can enter the human body through inhalation, ingestion, or skin contact. In occupational settings, inhalation and dermal exposure are the primary routes of contact with hazardous substances. Breathing in airborne chemicals such as gases, vapors, and particulate matter (including dust, smoke, fumes, aerosols, and mists) can lead to rapid absorption,

particularly when these substances reach the lungs. From there, they can quickly enter the bloodstream and travel to other organs.

While absorption through the skin is generally slower than through inhalation or ingestion, it can occur more rapidly if the skin is broken or damaged. Many chemical compounds, especially those that are fat-soluble, can pass through intact skin. Dermal exposure can lead to occupational skin diseases and, in some cases, systemic toxicity—especially with substances like pesticides and organic solvents. Workers may not always be aware that chemicals are being absorbed through their skin, making this type of exposure particularly concerning[2].

1.3. Classification hazards of chemicals agents:

A hazard is anything with the potential to cause bodily injury, and includes any physical, chemical, biological, mechanical, electrical, or ergonomic hazard. The Globally Harmonized System (GHS) divides hazardous chemicals in the workplace into different categories: physical hazards, health hazards, and environmental hazards (GHS 2007) [3].

Table I.1 : Globally Harmonized System hazard classes [4].

Hazard types	**Hazard categories**
Physical hazard	Explosives; Flammable gases; Flammable aerosols; Oxidizing gases; Gases under pressure; Flammable liquids; Flammable solids; Self-reactive substances; Pyrophoric liquids; Pyrophoric solids; Self-heating substances; Substances which on contact with: water, emit flammable gases; Oxidizing liquids; Oxidizing solids; Organic peroxides Corrosive to metals.
Heath hazard	Acute toxicity (oral, dermal, and inhalation); Skin corrosion/irritation; Serious eye damage/eye irritation; Respiratory sensitizer; Skin sensitizer; Mutagenicity; Carcinogenicity; Toxic to reproduction Specific Target organ toxicity following single exposure; Specific Target organ toxicity following repeat exposure; e Aspiration hazard.
Environnemental hazard	Acute hazards to the aquatic environment; Chronic hazards to the aquatic environment

1.4. Types of occupationnal diseases:

Cancer: Occupational cancers develop after a long latent period. The time between first exposure to the carcinogen and presentation of cancer is usually more than 10–15 years. It can be even longer as in the case of asbestos-related mesothelioma, which can take 40–50 years to develop. Susceptibility to occupational carcinogens is higher when the exposure happens at a younger age, or if there are combined exposures such as smoking and asbestos.

More recently, cancer has been associated with a wide variety of occupational exposures, including asbestos (mesothelioma and lung cancer), benzene (leukemia and lymphoma), and vinyl chloride (angiosarcoma of the liver). Other well-recognized associations between occupational exposure and cancer include a link between arsenic and cancer of the skin, lung, and liver ether and at cell carcinoma of the lung [5].

Occupational Neurological and Psychological Disease:

Occupational exposure to neurotoxic substances has been clearly associated with a variety of neurologic and psychological disorders, including lead-induced encephalopathy and neuropathy, parkinsonism in workers exposed to manganese, and both acute and chronic neuropathies from organophosphorus pesticides. Peripheral neuropathy has also been observed in workers exposed to solvents such as n-hexane.

Additional cases include chronic encephalopathy and peripheral neuropathy in pesticide production workers exposed to chlordecone (Kepone), severe neurologic disease from exposure to leptophos (Phosvel), and cognitive impairment linked to long-term solvent exposure. Despite these well-documented associations, the actual proportion of neurologic and psychiatric illnesses attributable to workplace neurotoxins remains unknown. It is also unclear whether these recognized cases are isolated incidents or indicators of a broader, underdiagnosed public health issue. Effectively addressing occupational neurologic disease in the United States requires a comprehensive strategy that includes preventing toxic exposures in the workplace, conducting mandatory premarket toxicological testing of all new chemicals and technologies, and improving clinical awareness and diagnostic accuracy among healthcare professionals [5].

1.5. Principles prevention from occupational diseases:

Many of the ODs that are the result of specific chemical agents are preventable. Three levels of prevention can be implemented. Primary prevention aims to prevent the occurrence of a disease by eliminating the causal agent or preventing it from causing bodily damage. Secondary prevention aims to detect disease in its early stages and to halt the progression of the

disease before it manifests as clinical symptoms and signs. Tertiary prevention is applicable to people with established disease, who require treatment and rehabilitation to minimize complications and disabilities or to improve quality of life if the disease is incurable. Once an OD due to chemical exposure is diagnosed, management of the disorder should go beyond prescribing medication to cure the condition. The following measures may be needed: suspension of the worker from further exposure; investigating and controlling the source of exposure; notification to relevant authorities; educating the patient and employer; identifying other workers who could also be exposed; rehabilitation; assessment of permanent disability; and compensation for affected workers[6].

The well know of the occupational exposures limits is considered as the perfect way to prevent a large amount of the occupational diseases.

2.Occupational exposure limits:

2.1. History:

Over the past 60 years, many organizations in numerous countries have proposed occupational exposure limits (OELs) for airborne contaminants. The limits or guidelines that have been the most widely accepted both in the United States and in most other countries are those issued annually by the American Conference of Governmental Industrial Hygienists (ACGIH) and are termed Threshold Limit Values (TLVs). The usefulness of establishing OELs for potentially harmful agents in the working environment has been demonstrated repeatedly since their inception .It has been claimed that whenever these limits have been implemented in a particular industry, no worker has been shown to have sustained serious adverse effects on his health as a result of exposure to these concentrations of an industrial chemical Although this statement is arguable with respect to the acceptability of OELs for those chemicals established before 1980, and later found to be carcinogenic, there is little doubt that millions of persons have avoided serious effects of workplace exposure due to their existence [7].

2.2. Intented use of OELs:

The ACGIH TLVs and most other OELs used in the United States, as well as most other countries, are limits that refer to airborne concentrations of substances and represent conditions under which “it is believed that nearly all workers may be repeatedly exposed day-after-day without adverse health effects.” (ACGIH, 2009). In some countries, the OEL is set at a concentration that attempts to protect virtually everyone. It is important to recognize that unlike some exposure limits for ambient air pollutants, contaminated water, or food additives set by

other professional groups or regulatory agencies, exposure to the TLV will not necessarily prevent discomfort or injury for everyone who is exposed. The ACGIH recognized long ago that because of the wide range in individual susceptibility, a small percentage of workers may experience discomfort from some substances at concentrations at or below the threshold limit and that a smaller percentage may be affected more seriously by aggravation of a preexisting condition or by development of an occupational illness [8].

2.3. Definitions:

A. Occupational Exposure Limits (OELS):

Occupational Exposure Limits (OELs) are vital tools for evaluating and tracking workers' exposure to hazardous substances. Used for decades in industrialized nations, they play a key role in preventing harmful health effects from exposure to chemical agents in the workplace. An OEL is defined as the maximum concentration of a substance, typically in the air, to which workers can be exposed repeatedly throughout their careers or briefly in acute situations, without causing harm to their health or that of future generations. By comparing the concentration of a hazardous substance in the air to its OEL, employers can assess the risks to workers and implement appropriate risk management strategies. Additionally, OELs help monitor and improve the effectiveness of existing safety measures [9].

B. Time-Weighted Average (TWA):

The time-weighted average concentration for a conventional 8-hour workday and 40-hour workweek exposure to a substance, to which it is believed that nearly all workers may be repeatedly exposed, day after day, without adverse health eff ects. The data are given after American Conference of Governmental Industrial Hygienists, ACGIH, National Institute for Occupational Safety and Health, NIOSH, and Occupational Safety & Health Administration, OSHA[10].

C. Short-Term Exposure Limit (STEL):

A Short-Term Exposure Limit (STEL) is defi ned by ACGIH as the concentration to which workers can be exposed continuously for a short period of time without suffering from irritation, chronic or irreversible tissue damage, or narcosis of sufficient degree to increase the likelihood of accidental injury, impair self-rescue or materially reduce work efficiency[11].

2.4. Approach of setting OELs:

The historical approach used to set OELs was based on human experience in the workplace. If airborne levels of a chemical were causing adverse health effects, then these levels were

reduced to a level that did not produce adverse health effects. In the latter half of the 1900s, as laboratory animal testing for the toxicity of chemicals became more common and more epidemiologic studies were done in workplaces, the approach used to set OELs was based on the "no-observed effect-level/safety factor" (NOEL/SF) approach. In this approach, all of the pertinent animal and human studies are reviewed and the highest dose that did not cause an effect in the most sensitive health end point (the NOEL) is identified. Once a NOEL has been identified, a set of uncertainty (or safety) factors are applied to this value to accommodate limitations in the data and to try to assume that workers are protected. The number and magnitude of these safety factors depend on the quality of the data. In general, some of these safety factors may include:

(1): a factor from 1 to 10 for animal-to-human (interspecies) extrapolation (if the NOEL is based on animal data)

(2): a factor from 1 to 10 for human-to-human intraspecies variability in response

(3): a factor from 1 to 10 to consider study duration (a long-term study being more helpful than a short-term study)

(4): a factor to consider the persistence of the drug in the body (or elimination half-life)

(5): a factor to accommodate for absorption efficiency by different routes of exposure.

If a NOEL is not available, then a lowest-observed-effect-level (LOEL) can be used The LOEL is the lowest level that causes an effect in the most sensitive end point. A safety factor from 1 to 10 may be considered for extrapolating a LOEL to a NOEL.

An equation that is used to determine an OEL for a chemical can be represented as follows5:

$$\boldsymbol{OELs} = [(\boldsymbol{NOEL}) \times (\boldsymbol{human\ body\ weight})] / [(\boldsymbol{safety\ factor}) \times (\boldsymbol{human\ breathing\ rate})]$$

NOEL: in milligram of chemical administered/kilogram of animal body weight/day.

Human body weight typically: is assumed to be 70 kiligrams for an adult male.

Safety factors: are numeric values for accomodating limitations in the data, as described above.

Breathing rate: in workers typically is assumed to be 10 m3 /8-hour workday[12].

2.5. Safety Data Sheet:

The safety data sheet (SDS), formerly known in the United States as the material safety data sheet (MSDS), contains important information for the safe handling of chemicals [13]. The Material Safety Data Sheet (MSDS) is a written document that identifies health and safety

information about any product containing one or more hazardous chemicals. This document provides vital information on chemical and physical hazards as well as information concerning safe use, handling, and storage. The MSDS can be extremely valuable in helping the emergency physician ascertain the potential toxicity of a given hazardous chemical substance. Emergency physicians should be aware of the existence of MSDS documents and be familiar with their format and the information available on them, but they should also be aware of the many limitations of the MSDS as a source of information [14].

CHAPTER II

GENERALITIES ABOUT ARTIFICIAL INTELLIGENCE AND CHEMOINFORMATICS

1.AI Methods:

1.1. Artificial intelligence:

Artificial intelligence (AI) has brought about a transformative shift across global industries, with particularly significant potential in healthcare. In the medical field, AI is used to analyze patient data—ranging from hospital visits and prescribed medications to lab tests and medical procedures—helping uncover insights that might remain hidden within vast datasets. By managing and interpreting large volumes of medical information, AI enables healthcare providers to make more informed decisions and improve patient outcomes. As a machine learning system, AI can analyze data in real time, enhancing the efficiency of research and clinical processes. These technologies are also instrumental in drug discovery and development, contributing to more effective healthcare management and patient treatment strategies. AI is especially beneficial in the management of diseases such as neurological disorders, cancer, diabetes, and cardiovascular conditions. Moreover, the more data AI systems process, the more intelligent and accurate they become, further advancing the pharmaceutical and healthcare sectors[15].

1.2. Machine Learning ML:

Machine learning (ML) is known as one of the most promising AI technologies, which is typically the study of computer algorithms that automate analytical model building. ML models are often made up of a set of rules, procedures, or sophisticated “transfer functions” that can be used to discover interesting data patterns or anticipate behavior. Machine learning is also known as predictive analytics that makes predictions about certain unknowns in the future through the use of data and is used to solve many real-world business issues, e.g., business risk prediction[16].

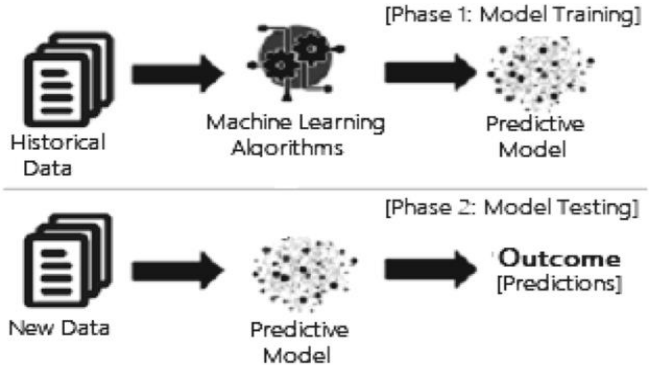


Figure II.1: A general structure of a machine learning system (training phase and testing phase)

1.3. Types of ML:

1.3.1. Supervised learning:

Supervised learning is applied when the data is in the form of input variables and output target values. The algorithm learns the mapping function from the input to the output. The availability of large scale labeled data samples makes it an expensive approach for tasks where data is scarce. These approaches can be broadly divided into two main categories [17]:

1.3.1.1. Classification:

Classifier algorithms are designed to determine whether a new data point belongs to one of several predefined classes. These classification models undergo a learning process using labeled examples from each category. During this training phase, the models identify patterns, correlations, and relationships within the data that enable them to distinguish one class from another. Once trained, the model can then accurately assign class labels to new, previously unseen data points based on the learned patterns. [18]

1.3.1.2. Regression:

Regression algorithms are used to predict a continuous output variable based on input features. These models analyze the relationships between the input variables and the target outcome to uncover underlying patterns. By learning these patterns during training, regression models can accurately estimate the values of new, unseen data points [19].

Table II.1 : types and algorithms of supervised learning.

Type	**Algorithm**
Classification	Logistic regression
	SVM (**Support Vector Machine**)
	Decision tree
	Random forest
	KNN (K-Nearest Neighbors)
Regression	Linear regression
	Polynomial regression
	Decision tree regression
	Random Forest regression
	SVR (**Support Vector Regression)**

1.3.1.3.SVR (support vector machine of regression):

Support vector regression (SVR) is a supervised machine learning technique to handle regression problems. SVR is useful because it balances model complexity and prediction error, and it has good performance for handling high-dimensional data. SVR is an extension to the Support Vector Machine (SVM) classification algorithm. SVR has additional advantages when compared to other regression methods. With the use of a kernel, SVR can provide an efficient way to handle a nonlinear regression problem by projecting the original feature into a kernel space where data can be linearly discriminated. Another benefit of SVR is that it learns a model to describe a variable's importance in characterizing the relationship between input and output, whereas in a traditional data regression method, one needs to assume a model that might not be accurate[20].

1.3.1.4. Kernel Method:

A. Definition:

Kernel Methods, a powerful class of algorithms for pattern analysis, have become a standard tool in data analysis, computational statistics, and machine learning applications due to their reliability, accuracy, and computational efficiency [10]. They have the capability to handle a very wide range of data types (e.g. sequences, vectors, networks) by working in a high dimensional feature space. This is obtained with the function Φ(x) which maps the data x from the original input space to the feature space. This embedding is performed by the 'kernel function' k (xk, xl), which efficiently computes the inner product hΦ(xk), Φ(xl)i between all pairs of data items xk and xl in the feature space. This results in the kernel matrix with the size determined by the number of data items. Any symmetric, positive semidefinite function is a valid kernel function (e.g. linear, polynomial, and diffusion kernels). They all correspond to a different transformation of the data, meaning that they extract a specific type of information from the data set[21].

1.3.2. Unsupervised learning:

Unsupervised learning is applied when the data is available only in the form of an input and there is no corresponding output variable. Such algorithms model the underlying patterns in the data in order to learn more about its characteristics.

One of the main types of unsupervised algorithms is **Clustering**. In this technique, inherent groups in the data are discovered and then used to predict output for unseen inputs [22].

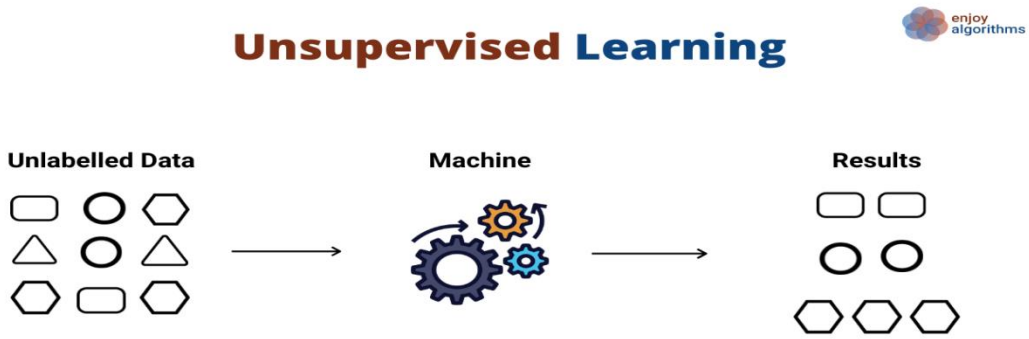


Figure II.2: Schema of unsupervised learning

1.3.3. Reinforcement Learning:

Reinforcement learning is another type of machine learning training strategy that rewards desired behaviors while punishing unwanted ones. A reinforcement learning agent, in general, is capable of perceiving and interpreting its surroundings, taking actions, and learning through trial and error, an environment-driven approachReinforcement learning is applied when the task at hand is to make a sequence of decisions towards a final reward[23].

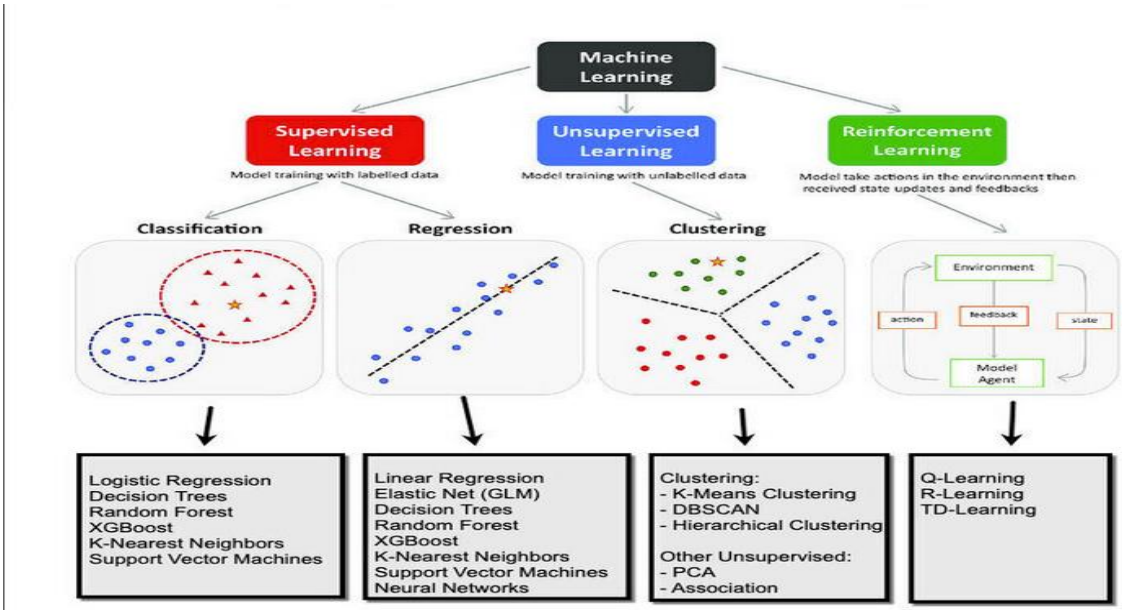


Figure II.3: Difference between reinforcement, supervised and unsupervised learning

1.3.4. Semi supervised learning:

To particular supervised and unsupervised tasks, semi-supervised learning can be regarded as a hybridization of both techniques explained above, as it uses both labeled and

unlabeled data to train a model. It could be efective for improving model performance when data must be labeled automatically without human interaction[24].

Machine learning approaches have the potential to contribute significantly to the development of effective models across various application domains, depending on their learning capabilities, the characteristics of the data, and the specific objectives of the task[25].

1.4. Deep learning DL:

Deep learning, is a set of algorithms and approaches inspired by how the human brain works. Deep learning architectures provide a lot of advantages for text classification, since they can perform extremely well with low-level engineering and computation. Deep learning (DL) algorithms require a large amount of training data than traditional ML algorithms. However, unlike standard machine learning algorithms (i.e. SVM and NB), DL classifiers do not have a learning threshold for the training data, therefore as more data are fed as more the algorithm will be well trained [26].

1.5. The Relation of AI with ML and DL:

Artificial intelligence (AI), machine learning (ML), and deep learning (DL) are three prominent terminologies used interchangeably nowadays to represent intelligent systems or software. The position of machine learning and deep learning within the artificial intelligence field is depicted in Fig. According to Fig, DL is a subset of ML which is also a subset of AI.

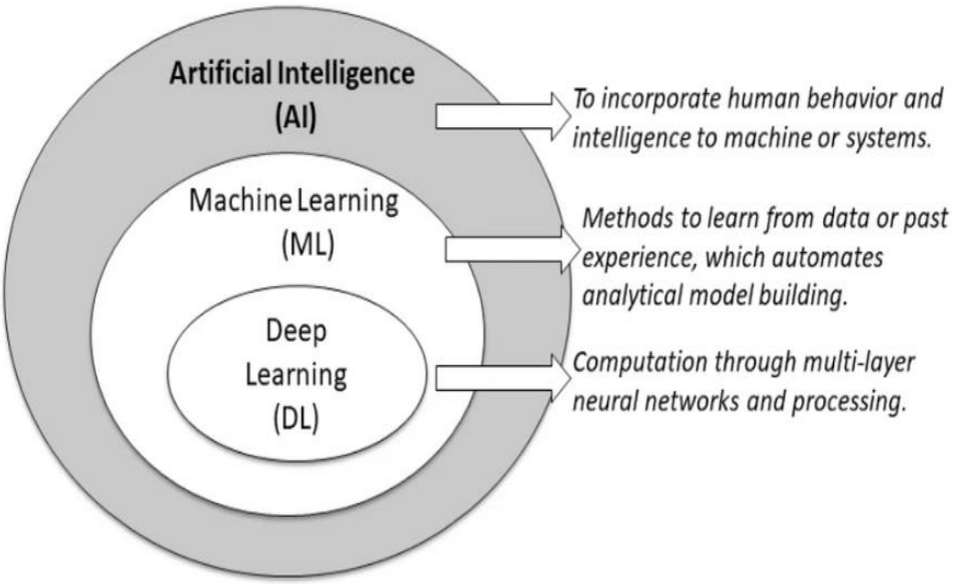


Figure II.4: Relation between AI, ML and DL

Artificial Intelligence (AI) broadly involves embedding human-like intelligence and behavior into machines or systems. Machine Learning (ML) is a subset of AI focused on enabling systems to learn from data or experience, thereby automating the creation of analytical models. Deep Learning (DL), another key subset, utilizes multi-layered neural networks to model and process data. The term "deep" signifies the multiple layers through which data passes to develop complex, data-driven models.

Both ML and DL are foundational technologies within AI, pushing the boundaries of what intelligent systems can achieve. They support the evolution of AI into what can be called "Smarter AI," where systems continuously improve through data-driven learning. These technologies are closely connected to the field of Data Science, which encompasses the entire process of extracting actionable insights from data within a specific context.

By learning from data, ML and DL contribute significantly to advanced analytics and intelligent decision-making, making them indispensable tools in data science. As such, they play a transformative role in today's world—powering sophisticated computational engines, enabling automation, and creating intelligent systems[27].

Over the past few years, deep learning techniques have been significantly developed and widely adopted for extracting information from various types of data. Depending on the specific characteristics of the input data, different deep learning architectures have been designed. Below, we highlight some of the main types:

2.Artificial neural networks ANNs

ANNs are artificial adaptive systems that are inspired by the functioning processes of the human brain. They are systems that are able to modify their internal structure in relation to a function objective. They are particularly suited for solving problems of the nonlinear type, being able to reconstruct the fuzzy rules that govern the optimal solution for these problems. The base elements of the ANN are the nodes, also called processing elements (PE), and the connections. Each node has its own input, from which it receives communications from other nodes and/or from the environment and its own output, from which it communicates with other nodes or with the environment. Finally, each node has a function f through which it transforms its own global input into output. Each connection is characterized by the strength with which pairs of nodes are excited or inhibited. Positive values indicate excitatory connections, the negative one's inhibitory connections[28].

2.1. Convolutional Neural Networks CNN:

The CNN is a kind of feedforward neural network that is able to extract features from data with convolution structures. Different from the traditional feature extraction methods, CNN does not need to extract features manually. The architecture of CNN is inspired by visual perception. A biological neuron corresponds to an artificial neuron; CNN kernels represent different receptors that can respond to various features; activation functions simulate the function that only neural electric signals exceeding a certain threshold can be transmitted to the next neuron. Loss functions and optimizers are something people invented to teach the whole CNN system to learn what we expect. CNN possesses many advantages:

- local connections—Each neuron is no longer connected to all neurons of the previous layer, but only to a small number of neurons, which is effective in reducing parameters and speed up convergence.
- Weight sharing—a group of connections can share the same weights, which reduces parameters further.
- Downsampling dimension reduction—a pooling layer harnesses the principle of image local correlation to downsample an image, which can reduce the amount of data while retaining useful information. It can also reduce the number of parameters by removing trivial features.

These three appealing characteristics make CNN one of the most representative algorithms in the deep learning field[29].

2.2. Recurrent Neural Network RNN:

Recurrent Neural Networks (RNNs) form an expressive model family for sequence tasks. They are powerful because they have a high-dimensional hidden state with nonlinear dynamics that enable them to remember and process past information. Furthermore, the gradients of the RNN are cheap to compute with backpropagation through time. A Recurrent Neural Network is a straightforward adaptation of the standard feed-forward neural network to allow it to model sequential data. At each time step, the RNN receives an input, updates its hidden state, and makes a prediction. The RNN's high dimensional hidden state and nonlinear evolution endow it with great expressive power, enabling the hidden state of the RNN to integrate information over many time steps and use it to make accurate predictions. Even if the non-linearity used by each unit is quite simple, iterating it over time leads to very rich dynamics [30].

2.3. Feed-Forword Neural Network FNN:

A feed-forward neural network (FNN) is an artificial neural network where connections between the nodes do not form a directed cycle. These artificial networks have a number of properties which make them particularly suited to complex pattern classification problems. On the other side, the success of their application to some real-world problems depends on a training algorithm. They need a training algorithm which reliably finds a nearly globally optimal set of weights in a relatively short time. Traditional algorithm for training FNN, called backpropagation, can often find a good set of weights in a reasonable amount of time. Backpropagation is a variation of gradient search and the key to backpropagation is a method for calculating the gradient of the error with respect to the weights for a given input by propagating error backwards through the network[31].

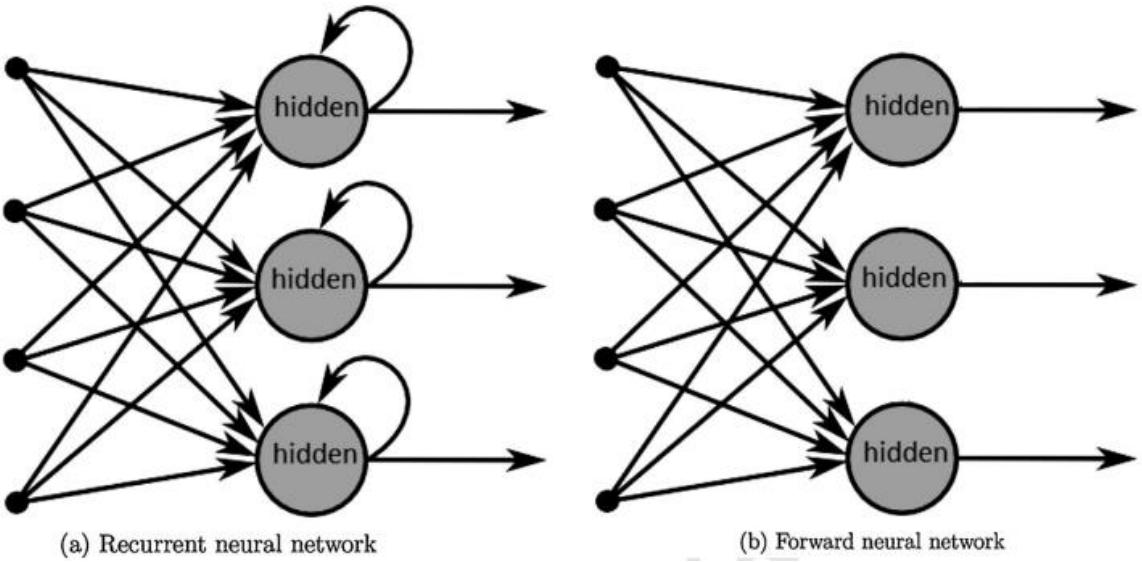


Figure II.5: Recurrent versus feedforward neural network.

2.3.1. Multi-layer perceptron's definition:

The MLP is a very simple model of biological neural networks and is based on the principle of a feedforward flow of information, i.e., the network is structured in a hierarchical way. The MLP consists of different layers where the information flows only from one layer to the next layer. Layers between the input and the output layers are called hidden layers. From a theoretical point of view, it is not necessary to consider more than one output unit because two or more output units could be realized by considering two or more MLPs in parallel[32].

2.3.2. Simple explication of MLPs algorithm:

1. **Forword pass:** it called the prediction step

   Input layer: receives features x from data

   Hidden layers: consists of neurons that perform **weighted summation** of the inputs they receive.

$$hi = f(\sum wij * xj + bi)$$

   where:

   - hi is the output of the hidden neuron i.
   - wij is the weight from input xj to neuron i.
   - bi is the bias term of the neuron.
   - $f()$ is the activation function (e.g., ReLU, Sigmoid).

If:

$(\sum wij * xj+bi)$ is positive (+) the neuron will activate and pass it forword. Otherwise, the output will be zero 0.

   Output layers: the output layers give the final prediction or decision.

$$\hat{y} = f(i\sum woi \cdot hi + bo)$$

   where:

   - $\hat{y}$ is the predicted output.
   - woi are the weights from the hidden layer hi to the output neuron.
   - bo is the bias of the output layer.

2. **Loss function:**

   The predicted output $\hat{y}$ is compared to the actual target value y using a **loss function** (e.g., Mean Squared Error for regression, Cross-Entropy for classification).
   The loss function quantifies how far off the prediction is from the actual target.

3. **Back propagation:** it called learning step

   Calculate the gradient of the loss with respect to each weight by **applying the chain rule** of calculus. This tells us how much each weight contributed to the error.

4. **Weighted update:**

   The weights are updated using the **gradients descent** and a **learning rate**

5. **Repeat the process:**

   The forward pass and backpropagation steps are repeated for many iterations (epochs) until the model converges to a minimum loss and the weights are optimized.

Activation Function:

Activation's function is needed for hidden layer of the NN to introduce nonlinearity. Without them NN would be same as plain perceptions. If linear function were used, NN would not be as powerful as they are. Activation function can be linear, threshold or sigmoid function. Sigmoid activation function is usually used for hidden layer because it combines nearly linear behavior, curvilinear behavior and nearly constant behavior depending on the input value Larose [2005]. To explain activation function figure will be used.

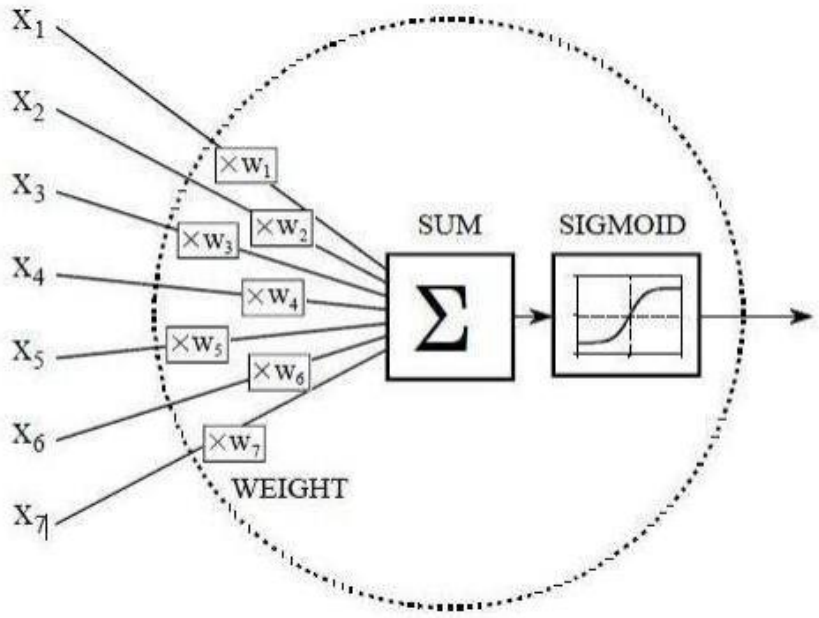


Figure II.6: Activation function

SUM is collection of the output nodes from hidden layer that have been multiplied by connection weights, added to get single number and put through sigmoid function (activation function). Input to sigmoid is any value between negative infinity and positive infinity number while the output can only be a number between 0 and 1 [33].

2.4. Back propagation:

2.4.1. Definition:

Back Propagation (BP) refers to a broad family of Artificial Neural Networks (ANN), whose architecture consists of different interconnected layers. The BP ANNs represents a kind of ANN, whose learning's algorithm is based on the Deepest-Descent technique. If provided with an appropriate number of Hidden units, they will also be able to minimize the error of nonlinear functions of high complexity. Theoretically, a BP provided with a simple layer of Hidden units is sufficient to map any functiony =f(x). Practically, it is often necessary to provide these ANNs with at least 2 layers of Hidden units, when the function to compute is particularly complex, or when the chosen data, in order to train the BP, are not particularly reliable, and a level filter is necessary on the features of Input. The BP are networks, whose learning's function tends to "distribute itself' on the connections, just for the specific correction algorithm of the weights that is utilized. This means that, in the case of BP, provided with at least a layer of Hidden units. [34]

2.4.2. Algorithm:

One of the most popular ANN algorithms is back propagation algorithm. Rojas [2005] claimed that BP algorithm could be broken down to four main steps. After choosing the weights of the network randomly, the back propagation algorithm is used to compute the necessary corrections. The algorithm can be decomposed in the following four steps:

- Feed-forward computation
- Back propagation to the output layer
- Back propagation to the hidden layer
- Weight updates

The algorithm is stopped when the value of the error function has become sufficiently small. This is very rough and basic formula for BP algorithm. There are some variations proposed by other scientist but Rojas definition seem to be quite accurate and easy to follow. The last step, weight updates is happening throughout the algorithm [34].

2.5.1 Gradient descent:

Gradient Descent is the core of the learning process in a Multilayer Perceptron (MLP).

It is the **main optimization tool** that tells the MLP how to **adjust its weights** in order to reduce the **error** between the predicted output and the actual output.

2.5.2. Importance of Gradient descent:

Reduces Error: Minimizes the loss function to improve prediction accuracy.

Enables Learning: Guides how weights should change to reduce mistakes.

Works with Backpropagation: Uses gradients to update weights effectively.

Handles Complexity: Finds good solutions in deep networks where no exact formula exists.

Efficient & Adaptable: Supports large datasets and various neural network types with optimized versions like Adam.

3.Cheminformatics

3.1. History:

Cheminformatics, also known as chemoinformatics, is a multidisciplinary domain combining chemistry, computer science, and information technology to enable the systematic handling of chemical information. This includes molecular structures, formulas, physicochemical properties, spectral data, and biological activities. Although the term was formally introduced by Frank K. Brown in 1998, the discipline's foundations were laid decades earlier. In 1957, Ray and Kirsch developed the first substructure search algorithm, initiating computational methods for structure-based queries. Hansch's introduction of Quantitative Structure–Activity Relationships (QSARs) in 1962 provided a mathematical basis for linking molecular features to biological activity, using variables such as hydrophobicity and electronic effects. In 1963, Vladutz's proposal to use computers for indexing chemical reactions further advanced the field by enabling computational retrosynthetic analysis.

Despite these early innovations, cheminformatics remained relatively under recognized until the emergence of high-throughput screening and the need to manage expansive compound libraries in pharmaceutical research. It has since become essential in various applications, including ligand-based drug design, virtual screening, predictive modeling, and molecular dynamics. Beyond drug development, its scope now encompasses fields like chemical genomics, systems biology, metabolomics, and materials science. The integration of artificial intelligence, increased computational resources, and vast chemical datasets continues to drive the evolution of cheminformatics, positioning it as a cornerstone of modern scientific inquiry[35].

3.2. Some definitions of cheminformatics:

- "The mixing of information resources to transform data into information, and information into knowledge, for the intended purpose of making better decisions faster in the arena of drug lead identification and optimization." Frank K. Brown, 1998.
- "[Chemoinformatics involves] ... the computer manipulation of two- or three-dimensional chemical structures and excludes textual information. This distinguishes the term from chemical information, largely a discipline of chemical librarians and does not include the development of computational methods." Peter Willett, 2002.

- ". . . the application of informatics to solve chemical problems." and ". . . chemoinformatics makes the point that you're using one scientific discipline to understand another scientific discipline." Johann Gasteiger, 2002
- "The set of approaches to computer-aided drug design that do not rely on the 3D structure of the [protein] target" John Van Drie, 2011 (personal communication)[36].

3.3. Applications:

The range of applications of cheminformatics is rich indeed; any field of chemistry can profit from its methods. The following lists different areas of chemistry and indicates some typical applications of cheminformatics.

- Storing data generated through experiments or from molecular simulation Retrieval of chemical Structures from chemical database (Software libraries).
- Prediction of physical, chemical and biological properties of chemical compounds.
- Elucidation of the structure of a compound based on spectroscopic data.
- Structure, Substructure, Similarity and diversity searching from chemical database
- High Throughput Screening (HTS) is the integration of technologies (elaborate Ory automation, assay technology, micro plate-based instrumentation, etc.) to quickly screen chemical compounds in search of a desired activity.
- Docking - Interaction between two macromolecules.
- Drug Discovery.
- Molecular Science, Materials Science, Food Science (nutraceuticals), Atmospheric chemistry, Polymer chemistry, Textile Industry, Combinatorial organic synthesis (COS) [37].

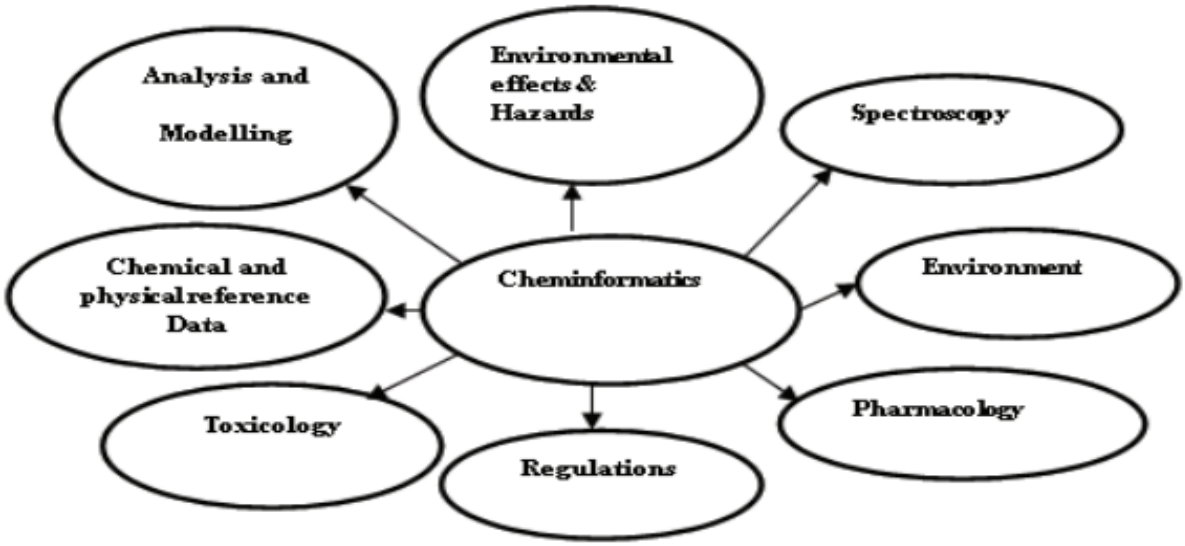


Figure II.7: Needs for Cheminformatics

3.4. Fingerprints:

Fingerprint representations of molecular structure and properties are a particularly complex form of descriptors. Fingerprints are typically encoded as binary bit strings whose settings produce, in different ways, a bit "pattern" characteristic of a given molecule. Fig. () illustrates the design of a model fingerprint. In this case, each bit accounts for a fraction of a value of a numerically encoded descriptor or, alternatively, a defined molecular fragment or structural key. Fingerprints are designed to account for different sets of molecular descriptors, structural fragments, possible connectivity pathways through a molecule, or different types of pharmacophores. If numerical descriptors are encoded in a binary format, their bit settings must account for specific value ranges. Thus, for each descriptor, meaningful value ranges should be determined prior to design of fingerprints, for example, by calculating descriptor values for large compound collections and surveying their distribution in histograms. In this way, appropriate bit settings can be defined that cover the distributions of numerical descriptors in large compound databases, as illustrated in Fig. (). By contrast, binary encoding of structural fragment-type descriptors is straightforward, since only their presence (i.e., 1) or absence (0) must be detected [38].

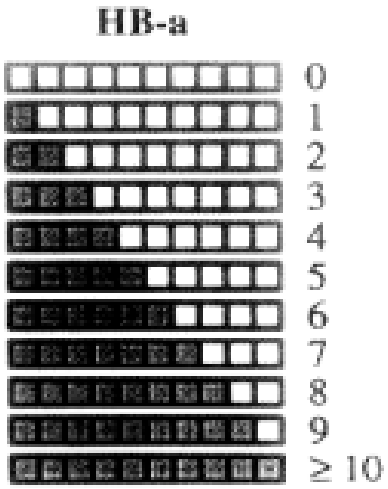


Figure II.8 : fingerprint of HB-a

Explication: Schematic representation of a binary molecular fingerprint. Values range of numerical descriptors "HB-a" (number of hydrogen bond acceptors) are encoded as indicated by shading. Non-shaded bit positions are set off (0). Gray shading means that a bit position is set on (1). For example, the value range of HB-a is encoded using a segment consisting of 10 bits. If all 10 bits are set to "0", no hydrogen bond acceptor is present. If the first five bits are set to "1", then the test molecule has five hydrogen bond acceptors.

The size of fingerprints can range from less than 100 to several thousands of bit positions. In "hashed" fingerprints, this is no longer the case because different features and/or values are mapped to overlapping big segments.

3.5. Descriptors and their classification:

Molecular descriptors (MDs) are numerical values associated with the structural features of compounds from databases, and they are necessary to apply chemoinformatics methods and relate their structures to a property or activity. A wide variety of MDs have been described in scientific literature until now. MDs are often classified according to the nature of the features they involve and their increasing complexity. According to this scheme, they can be referred as:

- One-dimensional (1D) descriptors, also known as "constitutional" descriptors, which encode chemical composition (counts of atoms, bonds, rings…), pharmacophore features (number of hydrogen-bond donors, acceptors, hydrophobic atoms…) or physicochemical properties (molecular weight, van der Waals volumes, Log P…). These

MDs are often employed for the prediction of simple physical properties and similarity/dissimilarity between compounds.

- Two-dimensional (2D) descriptors, encoding structural topology features such as size, branching, shape, etc. The most significant are the topological indices (TIs), which are determined from the application of graph theory [45] to chemistry. In this way, the molecules are considered as graphs with the atoms situated at the vertices and bonds represented by the edges. The connections between atoms can be described by various types of topological matrices (e.g. distance or adjacency matrices), which can be mathematically manipulated so as to derive a single number, usually known as graph invariant, graph theoretical index or topological index (TI).
- Three-dimensional (3D) descriptors, encoding three-dimensional shape and functionalities. These MDs depend of the spatial positions of the atoms, and their values vary for the same molecule depending upon the conformer chosen. From a structural point of view these descriptors can be considered such as more "realistic", since shape and functionality have definitively a crucial role in the recognition of small ligands by macromolecules. In opposite, they present the disadvantage that frequently no information about the active conformation is known, thus descriptors are obtained from an average of multiple conformers, that the representation can be biased.

A question that arises from the above list is to know which group of descriptors are the best to obtain some kind of knowledge through chemoinformatics applications. The main advantage of 1D/2D descriptors is their simplicity and rapidity of calculation, as well as the fact that conformation and alignment issues -sometimes related to 3D descriptors are completely avoided, and thus results obtained with them are globally more reproducible. Rather, 3D descriptors could be considered closest to "nature reality", since obviously molecules interact with their corresponding receptors in a three-dimensional way [39].

3.6. Quantitative Structure-Activity Relationship (QSAR):

The QSAR approach can be generally described as an application of data analysis methods and statistics to developing models that could accurately predict biological activities or properties of compounds based on their structures. Any QSAR method can be generally defined as an application of mathematical and statistical methods to the problem of finding empirical relationships (QSAR models) of the form Pi=k'(D1, D2,…,Dn), where Pi are biological activities (or other properties of interest) of molecules, D1, D2,…,Dn are calculated (or,

sometimes, experimentally measured) structural properties (molecular descriptors) of compounds, and k' is some empirically established mathematical transformation that should be applied to descriptors to calculate the property values for all molecules . The goal of QSAR modeling is to establish a trend in the descriptor values, which parallels the trend in biological activity [40].

Experimental part

CHAPTER III

AI MODELING OF OCCUPATIONAL EXPOSURE LIMITS FOR PHARMACEUTICAL COMPOUNDS USING CHEMOINFORMATICS

1.Introduction

In the pharmaceutical and chemical manufacturing sectors, workers are frequently exposed to a wide range of bioactive substances. Many of these compounds, such as highly potent active pharmaceutical ingredients (HPAPIs), pose substantial health risks even at trace levels. Ensuring worker safety, therefore, requires robust and proactive exposure control strategies. Among these, the **Occupational Exposure Banding (OEB)** system has emerged as a vital tool for early hazard assessment and risk mitigation. [41].

The increasing diversity of chemical substances in pharmaceutical production has far outpaced the development of official OELs. The U.S. EPA reports that over 85,000 chemical substances are in commercial use, yet fewer than 1,000 have assigned OELs [41].

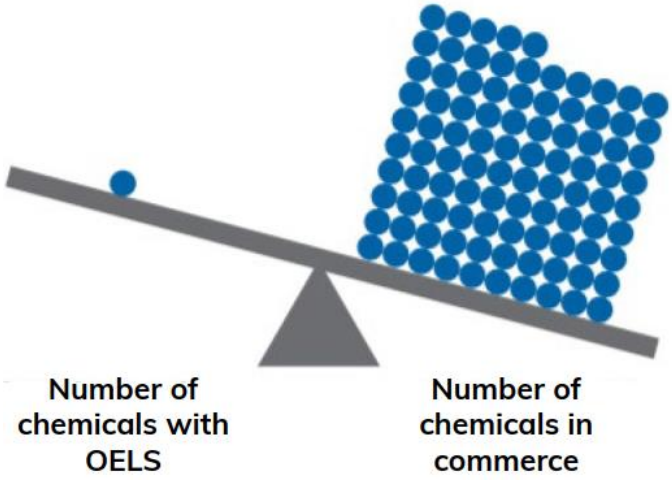


Figure III .1 There are many more chemicals on the market than have OELs.

Occupational Exposure Banding (also known as hazard banding or health hazard banding) is a structured process that classifies chemical substances into discrete hazard categories based on their potential for adverse health effects. This classification facilitates the determination of appropriate handling and containment measures, especially in cases where detailed toxicological data is limited or when no formal **Occupational Exposure Limit (OEL)** has yet been established.

Originally developed and promoted by agencies such as **NIOSH**, OEB systems offer a tiered and semi-quantitative approach that complements traditional OEL frameworks. While OELs are authoritative exposure thresholds expressed in quantitative terms (e.g., µg/m³), OEBs provide qualitative guidance, enabling pharmaceutical organizations to implement suitable safety protocols during early research and development stages or for new chemical entities [41].

With the emergence of artificial intelligence (AI) and chemoinformatics, predictive modeling has become a transformative approach in this area. By leveraging molecular descriptors and fingerprints derived from SMILES (Simplified Molecular Input Line Entry System) notations, machine learning models can estimate the likely OEB class of a compound based on its structure alone. This chapter

presents a hybrid AI framework that integrates convolutional neural networks (CNNs) for feature extraction with classical classifiers (e.g., SVM, MLP, XGBoost) for OEB prediction. The implementation is showcased through a custom-built **graphical user interface** application that enables real-time prediction and visualization from user-provided chemical inputs.

2.Theoretical Background

2.1. Occupational Exposure Limits (OELs)

OELs represent established, authoritative thresholds for airborne concentrations of hazardous substances in workplace air. These values, expressed typically in ppm or mg/m^3, are set by regulatory bodies such as the Occupational Safety and Health Administration (OSHA), the American Conference of Governmental Industrial Hygienists (ACGIH), and the European Medicines Agency (EMA). OELs are grounded in comprehensive toxicological and epidemiological data, considering both health risk and feasibility of implementation.

2.2. Occupational Exposure Bands (OEBs)

Occupational Exposure Bands (OEBs) are categorical indicators of the hazard level associated with exposure to a chemical substance in the workplace. They serve as a practical tool to estimate risk and guide control measures in environments where complete toxicological profiles are not available. The OEB system classifies substances into five or six levels (often labeled A to E or 1 to 5/6), with higher levels corresponding to greater toxicity and lower acceptable airborne concentrations.

The purpose of implementing OEBs includes

- **Protecting worker health** by minimizing chemical exposure
- **Facilitating containment decisions** during early development
- **Standardizing risk communication** between stakeholders
- **Supporting objective hazard assessments** when quantitative data is lacking

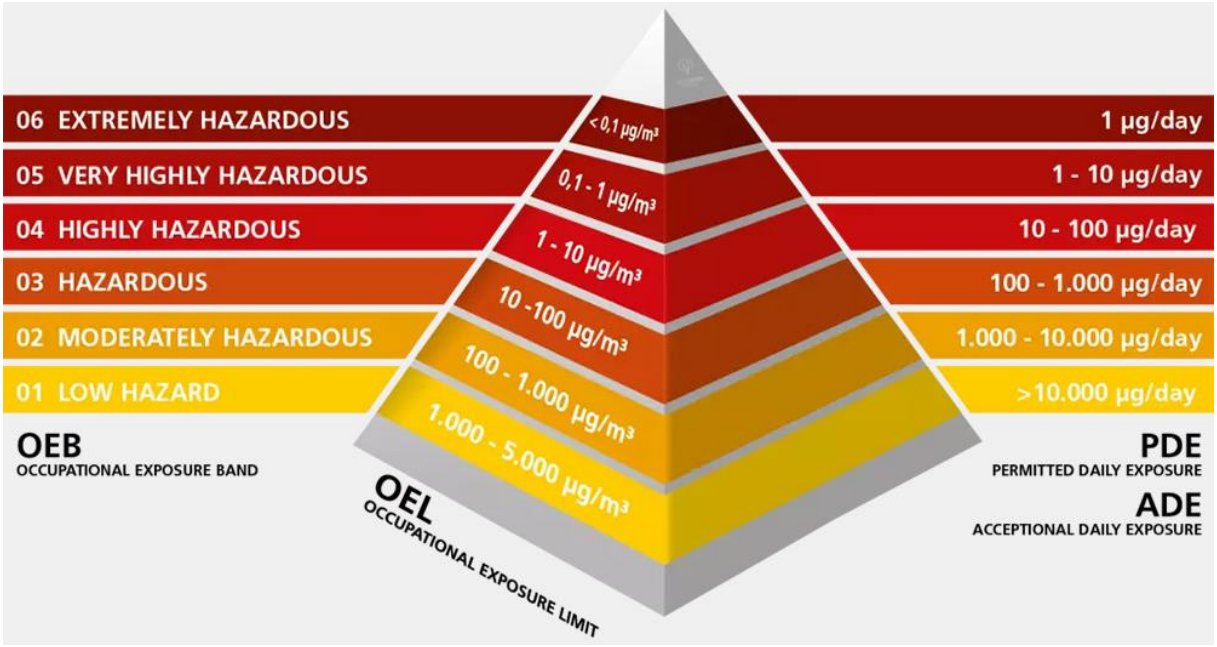


Figure III .2: OEB levels classification.

2.3.OEB Classification Arrangement

The classification into OEB levels is based on multiple toxicological endpoints, including:

- Eye and skin irritation or corrosion.
- Respiratory sensitization.
- Acute toxicity.
- Specific target organ toxicity.
- Reproductive and developmental toxicity.
- Genotoxicity.
- Carcinogenicity.

Each substance is scored across these domains, and an overall band from **A (least hazardous)** to **E (most hazardous)** is assigned. This grading correlates to specific **OEL ranges** and **containment strategies**, as shown in the table below:

Table III .1: Occupational exposure bands (OEB).

OEB Level	OEL Range (µg/m³)	Hazard Description	Recommended Containment
OEB 1 (A)	>1000	Non-hazardous	Local extraction, open handling
OEB 2 (B)	100–1000	Low hazard	Downflow booths, exhaust systems
OEB 3 (C)	10–100	Mildly hazardous	Glove bags, flexible isolators, continuous liners
OEB 4 (D)	1–10	Hazardous	RABS, isolators with glove ports
OEB 5 (E)	<1	Highly hazardous	Closed-system isolators, advanced filtration systems
OEB 6	<0.1	Extremely potent	Fully sealed isolators, robotic or aseptic systems

3.Methodology

3.1. Data collection and presentation

The study's dataset comprises 221 pharmaceutical active ingredients (APIs). These APIs were sourced from the official repository of the Algerian Ministry of Pharmaceutical Industry, which lists all substances approved for production or use within Algerian pharmaceutical manufacturing facilities. This foundational data was then integrated with information from Safety Data Sheets (SDS), published in accordance with United States Pharmacopeia (USP) standards, providing crucial details on physicochemical properties, toxicokinetic profiles, acute and chronic toxicity values, and recommended occupational exposure guidelines. This comprehensive approach ensured the dataset reflected both international toxicological standards and region-specific industrial relevance, creating a robust foundation for Occupational Exposure Band (OEB) modeling and AI-driven hazard classification. Figure **III .3** shows the OEB classes' distribution of data; in parallel, the molecules used are represented in Figure **III .4**.

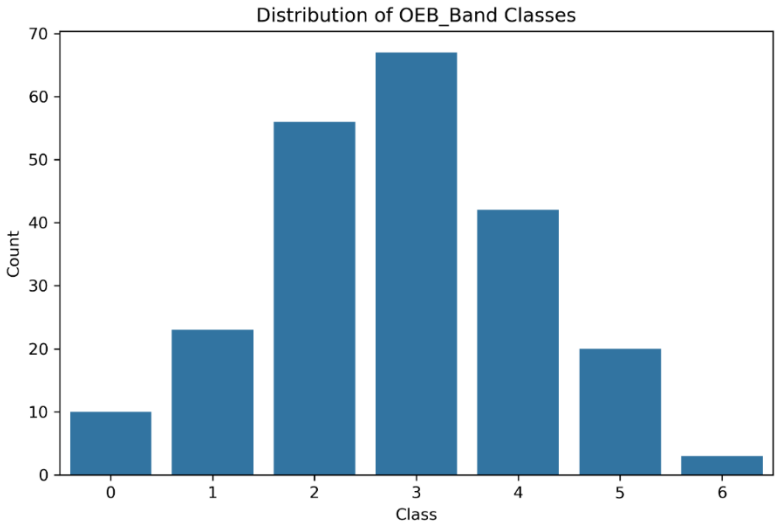
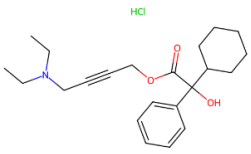
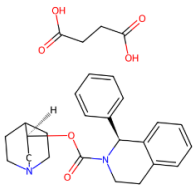


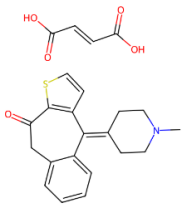
Figure III .3: Data OEB Classes Distribution.



OXYBUTYNINE CHLORHYDRATE
CAS: 1508-65-2



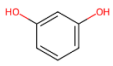
SOLIFENACINE SUCCINATE
CAS: 242478-38-2



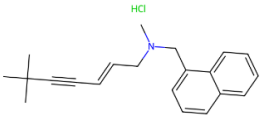
KETOTIFENE⁺¹ (SOLUS FORME DE FUMARATE)
CAS: 34580-14-8

OXYDE DE MAGNESIUM
CAS: 1309-48-4

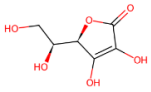
CALCIUM CARBONATE
CAS: 471-34-1



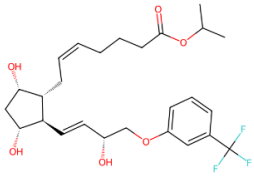
RESORCINOL
CAS: 108-46-3



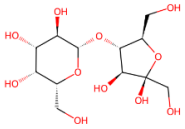
TERBINAFINE CHLORHYDRATE EXPRIME EN TERBINAFINE
CAS: 78628-80-5



ACIDE ASCORBIQUE
CAS: 50-81-7



TRAVOPROST
CAS: 157283-68-6

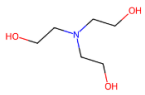


LACTULOSE
CAS: 4618-18-2

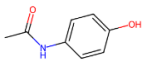
AMMONIUM C
CAS: 12125-02-9

OXYDE DE ZINC
CAS: 1314-13-2

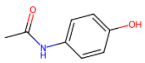
ALLOPURINOL
CAS: 315-30-0



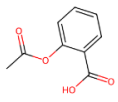
TROLAMINE PURE
CAS: 102-71-6



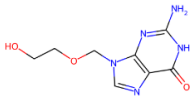
PARACETAMOL
CAS: 103-90-2



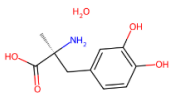
PARACETAMOL
CAS: 103-90-2



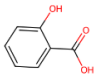
ACIDE ACETYLSALICYLIQUE
CAS: 50-78-2



ACYCLOVIR
CAS: 59277-89-3

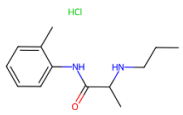


METHYLDOPA SESQUIHYDRATEE EXPRIME EN METHYLDOPA ANHYDRE
CAS: 41372-08-1

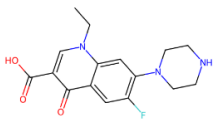


ACIDE SALICYLIQUE
CAS: 69-72-7

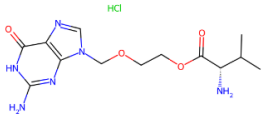
ACIDE SALICYLIQUE
CAS: 69-72-7



PRILOCAINE
CAS: 1786-81-8



NORFLOXACINE
CAS: 70458-96-7

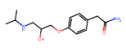


VALACICLOVIR CHLORHYDRATE
CAS: 124832-27-5

CAMPHOR
CAS: 464-49-3

IBUPROFENE
CAS: 15687-27-1

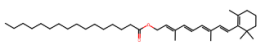
IBUPROFENE
CAS: 15687-27-1



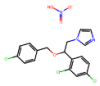
ATENOLOL
CAS: 29122-68-7

GABAPENTINE
CAS: 60142-96-3

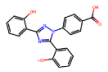
ENTACAPONE
CAS: 130929-57-6



RETINYL PALMITATE
CAS: 79-81-2



ECONAZOLE NITRATE
CAS: 24169-02-6

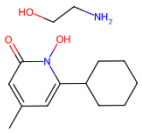


DEFERASIROX
CAS: 201530-41-8

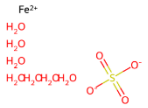
METFORMINE HCL
CAS: 1115-70-4

ACIDE VALPROIQUE (SOUS FORME DE VALPROATE DE SODIUM)
CAS: 99-66-1

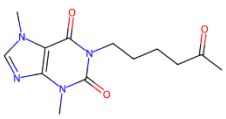
SULFAMETHOXAZOLE / TRIMETHOPRIME
CAS: 723-46-6



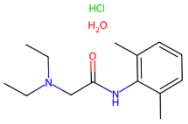
CICLOPIROX OLAMINE
CAS: 41621-49-2



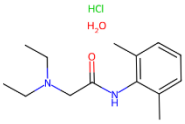
SULFATE FERREUX
CAS: 7782-63-0



PENTOXIFYLLINE
CAS: 167687

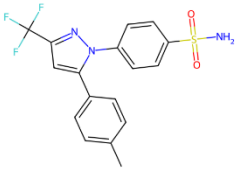


PHENAZONE/CHLORHYDRATE DE LIDOCAINE
CAS: 6108-05-0

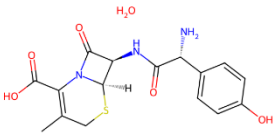


LIDOCAINE CHLORHYDRATE
CAS: 6108-05-0

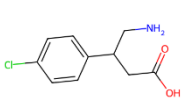
CIPROFLOXACINE
CAS: 85721-33-1



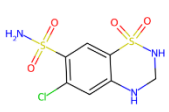
CELECOXIB
CAS: 169590-42-5



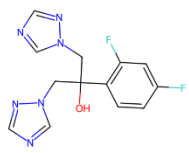
CEFADROXIL MONOHYDRATE EXPRIME EN CEFADROXIL ANHYDRE
CAS: 66592-87-8



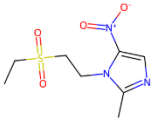
BACLOFENE
CAS: 1134-47-0



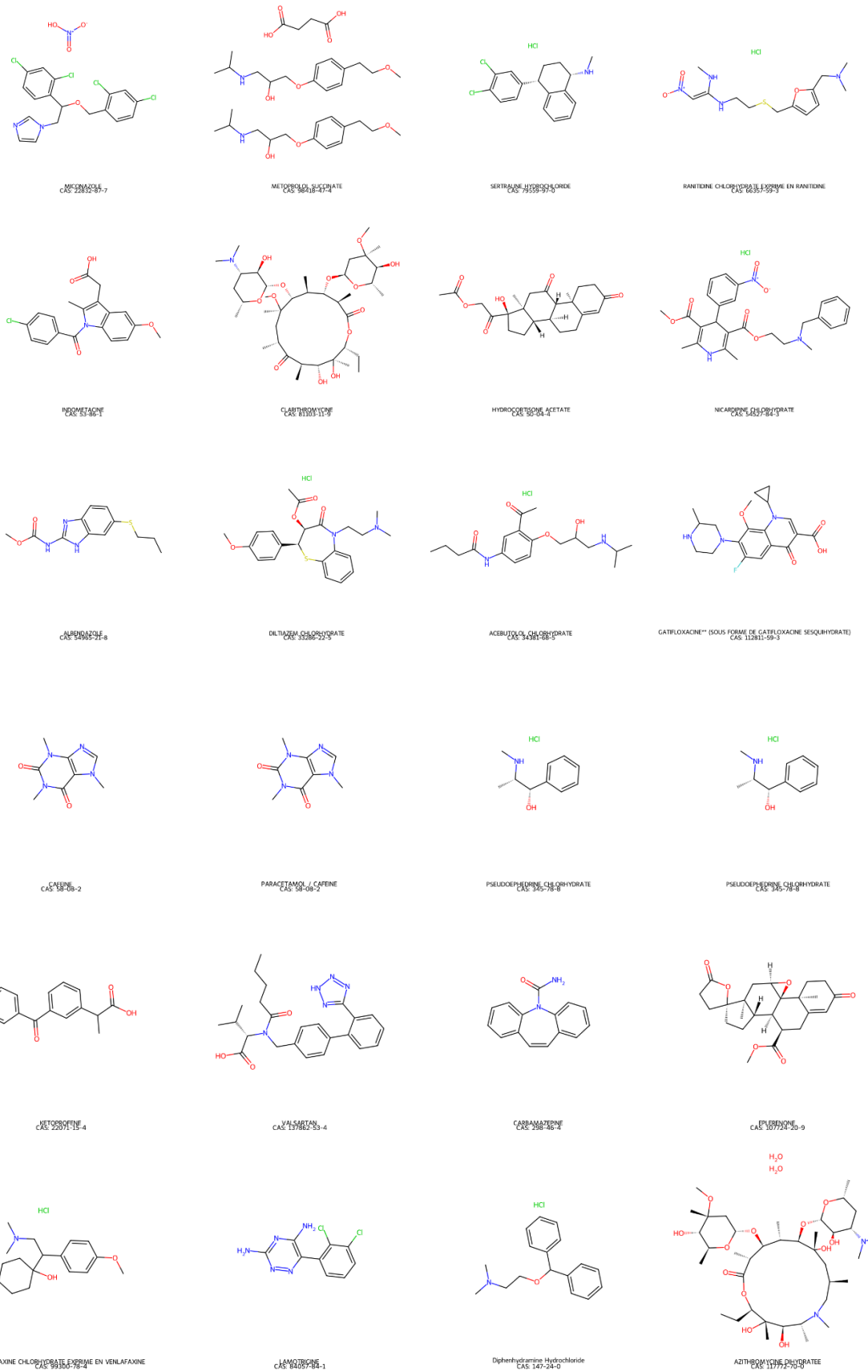
HYDROCHLOROTHIAZIDE
CAS: 58-93-5

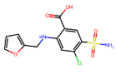


LUCONAZOLE
CAS: 88386-73-4



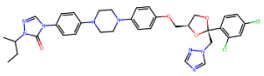
TINIDAZOLE
CAS: 19387-91-8


MICONAZOLE
CAS: 22832-87-7
METOPROLOL SUCCINATE
CAS: 98418-47-4
SERTRALINE HYDROCHLORIDE
CAS: 79559-97-0
RANITIDINE CHLORHYDRATE EXPRIME EN RANITIDINE
CAS: 66357-59-3
INDOMETACINE
CAS: 53-86-1
CLARITHROMYCINE
CAS: 81103-11-9
HYDROCORTISONE ACETATE
CAS: 50-04-4
NICARDIPINE CHLORHYDRATE
CAS: 54527-84-3
ALBENDAZOLE
CAS: 54965-21-8
DILTIAZEM CHLORHYDRATE
CAS: 33286-22-5
ACEBUTOLOL CHLORHYDRATE
CAS: 34381-68-5
CATIFLOXACINE** (SOUS FORME DE CATIFLOXACINE SESQUIHYDRATE)
CAS: 112811-59-3
CAFEINE
CAS: 58-08-2
PARACETAMOL / CAFEINE
CAS: 58-08-2
PSEUDOEPHEDRINE CHLORHYDRATE
CAS: 345-78-8
PSEUDOEPHEDRINE CHLORHYDRATE
CAS: 345-78-8
KETOPROFENE
CAS: 22071-15-4
VALSARTAN
CAS: 137862-53-4
CARBAMAZEPINE
CAS: 298-46-4
EPLERENONE
CAS: 107724-20-9
VENLAFAXINE CHLORHYDRATE EXPRIME EN VENLAFAXINE
CAS: 99300-78-4
LAMOTRIGINE
CAS: 84057-84-1
Diphenhydramine Hydrochloride
CAS: 147-24-0
AZITHROMYCINE DIHYDRATEE
CAS: 117772-70-0

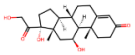


FUROSEMIDE
CAS: 54-31-9

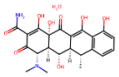
CLOMIPRAMINE CHLORHYDRATE
CAS: 17321-77-6



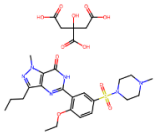
ITRACONAZOLE
CAS: 84625-61-6



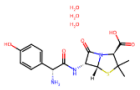
HYDROCORTISONE
CAS: 50-23-7



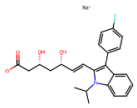
DOXYCYCLINE MONOHYDRATE EXPRIME EN DOXYCYCLINE
CAS: 17086-28-1



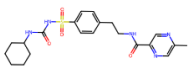
SILDENAFIL CITRATE
CAS: 171599-83-0



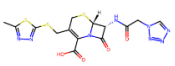
AMOXICILLINE TRIHYDRATE E EXPRIME EN AMOXICILLINE ANHYDRE
CAS: 61336-70-7



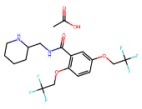
FLUVASTATINE SODIQUE EXPRIME EN FLUVASTATINE
CAS: 93957-55-2



CILOSTAZIDE
CAS: 73963-72-1

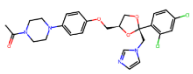


CEFAZOLINE
CAS: 25953-19-9

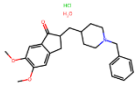


ACETATE DE FLECAINIDE
CAS: 54143-56-5

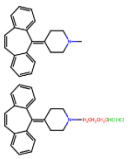
DESLORATADINE
CAS: 100643-71-8



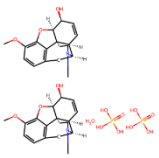
KETOCONAZOLE
CAS: 65277-42-1



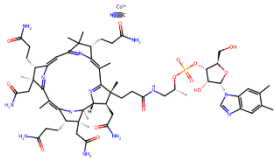
DONEPEZIL CHLORHYDRATE
CAS: 884740-09-4



CYPROHEPTADINE CHLORHYDRATE EXPRIME EN CYPROHEPTADINE
CAS: 41354-29-4



CODEINE PHOSPHATE HEMIHYDRATE
CAS: 41444-62-6



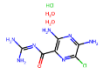
CYANOCOBALAMINE
CAS: 68-19-9

HYDROXYCARBAMIDE
CAS: 127-07-1

METRONIDAZOLE
CAS: 443-48-1

CAPTOPRIL
CAS: 62571-86-2

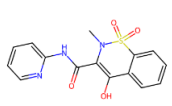
MINOXIDIL
CAS: 38304-91-5



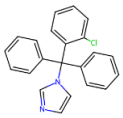
AMILORIDE (SOUS FORME DE CHLORHYDRATE)**/HYDROCHLOROTHIAZIDE
CAS: 17440-83-4

AMITRIPTYLINE CHLORHYDRATE EXPRIME EN AMITRIPTYLINE
CAS: 549-18-8

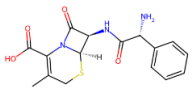
MEMANTINE CHLORHYDRATE
CAS: 41100-52-1



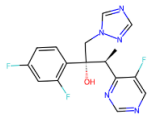
PIROXICAM
CAS: 36322-90-4



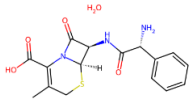
CLOTRIMAZOLE
CAS: 23593-75-1



CEFALEXINE MONOHYDRATE
CAS: 23325-78-2

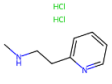


VORICONAZOLE
CAS: 137234-62-9

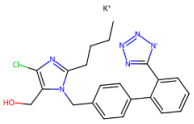


CEFALEXINE
CAS: 23325-78-2

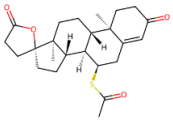
CEFACLOR
CAS: 70356-03-5



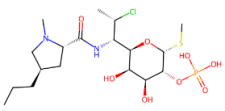
BETAHISTINE (SOUS FORME DE DICHLORHYDRATE)**
CAS: 5579-84-0



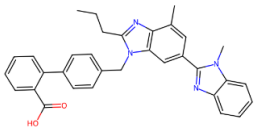
LOSARTAN POTASSIQUE
CAS: 124750-99-8



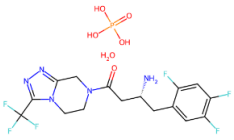
SPIRONOLACTONE / ALTIZIDE
CAS: 52-01-7



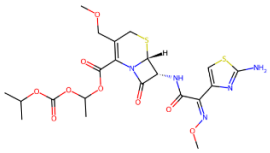
CLINDAMYCINE PHOSPHATE EXPRIME EN CLINDAMYCINE
CAS: 24729-96-2



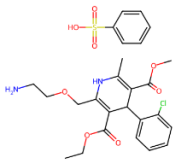
TELMISARTAN
CAS: 144701-48-4



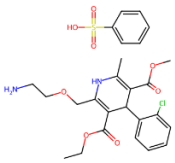
SITAGLIPTINE PHOSPHATE MONOHYDRATE EXPRIME EN SITAGLIPTINE
CAS: 654671-77-9



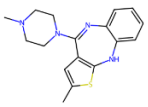
CEFPODOXIME** (SOUS FORME DE CEFPODOXIME PROXETIL)
CAS: 87239-81-4



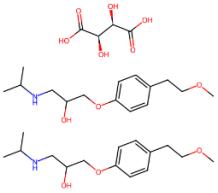
AMLODIPINE BESILATE
CAS: 111470-99-6



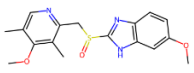
AMLODIPINE BESILATE
CAS: 111470-99-6



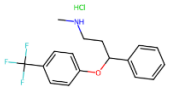
OLANZAPINE
CAS: 132539-06-1



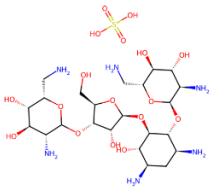
METOPROLOL
CAS: 56392-17-7



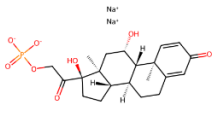
OMEPRAZOLE
CAS: 73590-58-6



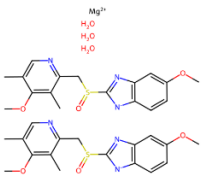
FLUOXETINE CHLORHYDRATE EXPRIME EN FLUOXETINE
CAS: 56296-78-7



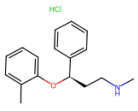
NEOMYCINE
CAS: 1405-10-3



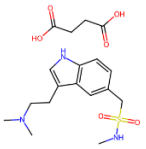
PREDNISOLONE SODIUM PHOSPHATE
CAS: 125-02-0



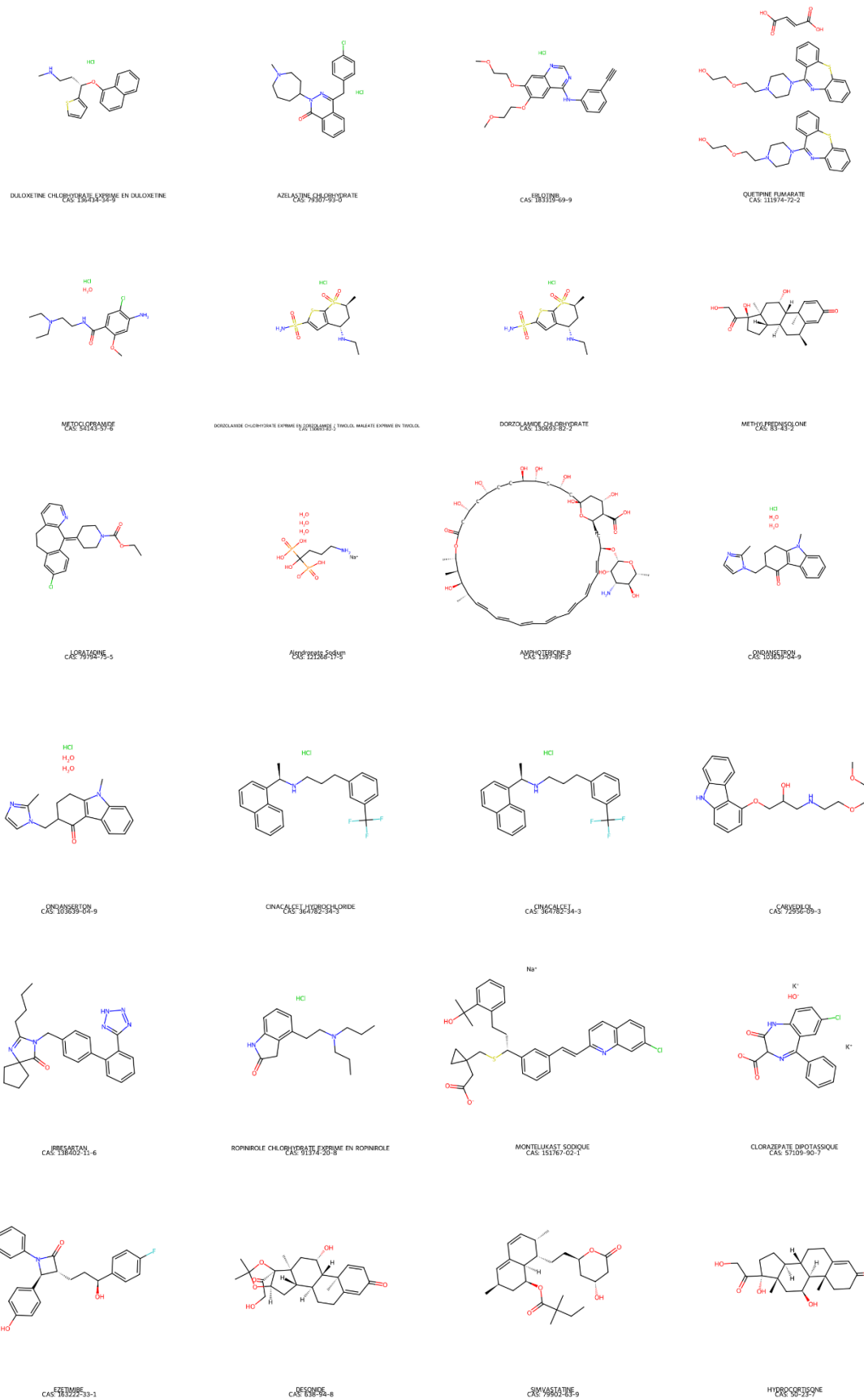
ESOMEPRAZOLE MAGNESIUM TRIHYDRATE
CAS: 217087-09-7

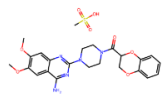


ATOMOXETINE HCL
CAS: 82248-59-7

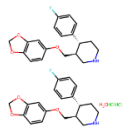


SUMATRIPTAN SUCCINATE
CAS: 103628-48-4


DULOXETINE CHLORHYDRATE EXPRIME EN DULOXETINE
CAS: 136434-34-9
AZELASTINE CHLORHYDRATE
CAS: 79307-93-0
ERLOTINIB
CAS: 183319-69-9
QUETIPINE FUMARATE
CAS: 111974-72-2
METOCLOPRAMIDE
CAS: 54143-57-6
DORZOLAMIDE CHLORHYDRATE EXPRIME EN DORZOLAMIDE (TIMOLOL MALEATE EXPRIME EN TIMOLOL
CAS: 130693-82-2
DORZOLAMIDE CHLORHYDRATE
CAS: 130693-82-2
METHYLPREDNISOLONE
CAS: 83-43-2
LORATADINE
CAS: 79794-75-5
Alendronate Sodium
CAS: 121268-17-5
AMPHOTERICINE B
CAS: 1397-89-3
ONDANSETRON
CAS: 103639-04-9
ONDANSERTON
CAS: 103639-04-9
CINACALCET HYDROCHLORIDE
CAS: 364782-34-3
CINACALCET
CAS: 364782-34-3
CARVEDILOL
CAS: 72956-09-3
IRBESARTAN
CAS: 138402-11-6
ROPINIROLE CHLORHYDRATE EXPRIME EN ROPINIROLE
CAS: 91374-20-8
MONTELUKAST SODIQUE
CAS: 151767-02-1
CLORAZEPATE DIPOTASSIQUE
CAS: 57109-90-7
EZETIMIBE
CAS: 163222-33-1
DESONIDE
CAS: 638-94-8
SIMVASTATINE
CAS: 79902-63-9
HYDROCORTISONE
CAS: 50-23-7



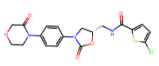
DOXAZOSINE MESILATE
CAS: 77883-43-3



PAROXETINE(SOUS FORME DE CHLORHYDRATE HEMIHYDRATEE)
CAS: 110429-35-1

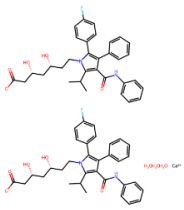
TADALAFIL
CAS: 171596-29-5

CLOPIDOGREL HYDROGENSULFATE EXPRIME EN CLOPIDOGREL
CAS: 120202-66-6

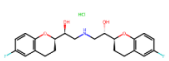


RIVAROXABAN
CAS: 366789-02-8

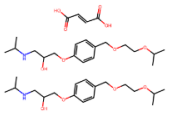
CHLORPROMAZINE HYDROCHLORIDE
CAS: 69-09-0



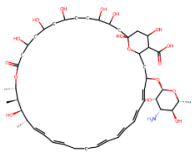
ATORVASTATINE CALCIQUE TRIHYDRATE EXPRIME EN ATORVASTATINE
CAS: 344423-98-9



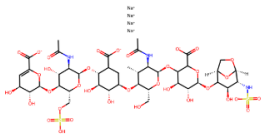
NEBIVOLOL CHLORHYDRATE EXPRIME EN NEBIVOLOL
CAS: 152520-56-4



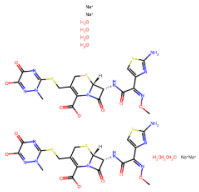
BISOPROLOL FUMARATE
CAS: 104344-23-2



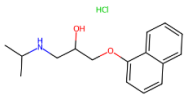
NYSTATINE
CAS: 1400-61-9



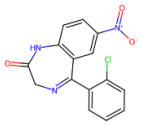
ENOXAPARINE SODIQUE
CAS: 679809-58-6



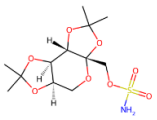
CEFTRIAXONE
CAS: 104376-79-6



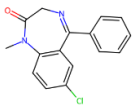
PROPRANOLOL
CAS: 318-98-9



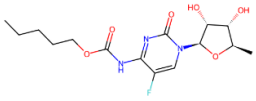
CLONAZEPAM
CAS: 1622-61-3



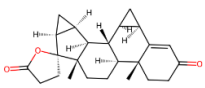
TOPIRAMATE
CAS: 97240-79-4



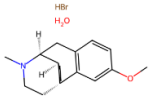
ERAZEPAM
CAS: 439-14-5



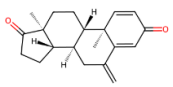
CAPECITABINE
CAS: 154361-50-9



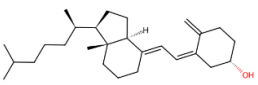
DROSPIRENONE
CAS: 67392-87-4



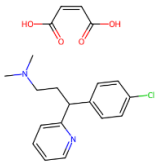
DEXTROMETHORPHANE HYDROBROMIDE
CAS: 6700-34-1



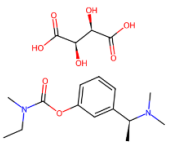
EXEMESTANE
CAS: 107868-30-4



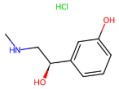
CHOLECALCIFEROL
CAS: 67-97-0



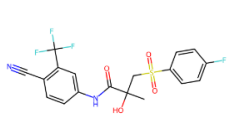
CHLORPHENIRAMINE MALEATE
CAS: 113-92-8



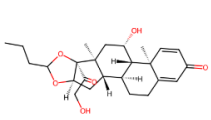
RIVASTIGMINE HYDROGENOTARTRATE EXPRIME EN RIVASTIGMINE
CAS: 129101-54-8



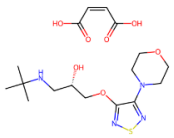
PHENYLEPHRINE CHLORHYDRATE
CAS: 61-76-7



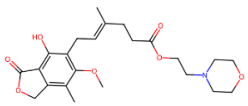
BICALUTAMIDE
CAS: 90357-06-5



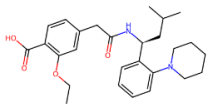
BUDESONIDE
CAS: 51333-22-3



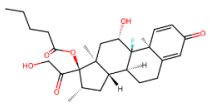
TIMOLOL
CAS: 26921-17-5



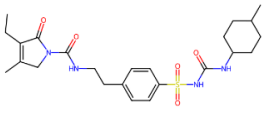
MYCOPHENOLATE MOFETIL
CAS: 128794-94-5



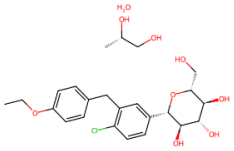
REPAGLINIDE
CAS: 135062-02-1



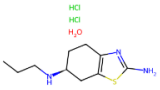
BETAMETHASONE VALERATE
CAS: 2152-44-5



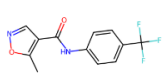
GLIMEPIRIDE
CAS: 93479-97-1



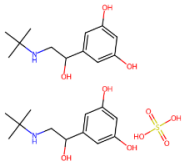
DAPAGLIFLOZINE PROPANEDIOL
CAS: 960404-48-2



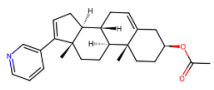
PRAMIPEXOLE DICHLORHYDRATE
CAS: 19121-81-9



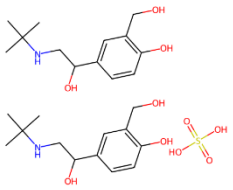
LEFLUNOMIDE
CAS: 75706-12-6



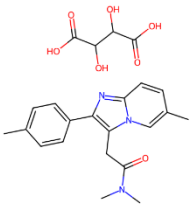
TERBUTALINE SULFATE
CAS: 23031-32-5



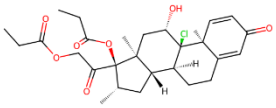
ABIRATERONE ACETATE
CAS: 154229-18-2



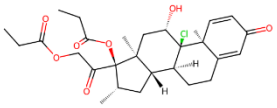
SALBUTAMOL SULFATE
CAS: 51022-70-9



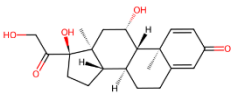
ZOLPIDEM TARTRATE
CAS: 99294-93-6



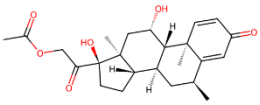
BECLOMETASONE DIPROPIONATE
CAS: 05534-09-8



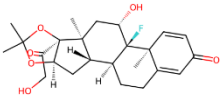
BECLOMETASONE
CAS: 5534 - 09 - 8



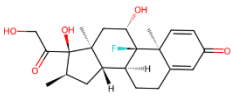
PREDNISOLONE
CAS: 50-24-8



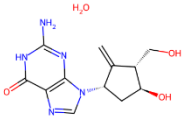
METHYLPREDNISOLONE ACETATE
CAS: 53-36-1



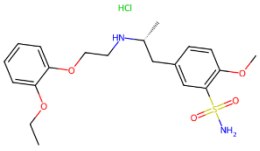
TRIAMCINOLONE ACETONIDE
CAS: 76-25-5



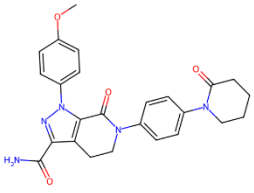
DEXAMETHASONE
CAS: 50-02-2



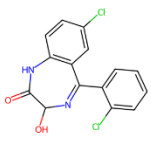
ENTECAVIR
CAS: 209216-23-9



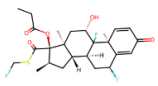
TAMSULOSINE CHLORHYDRATE
CAS: 106463-17-6

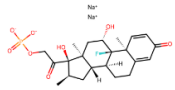


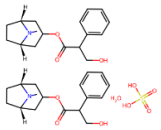
APIXABAN
CAS: 503612-47-3

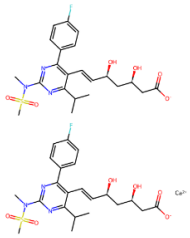


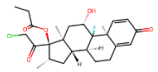
LORAZEPAM
CAS: 846-49-1

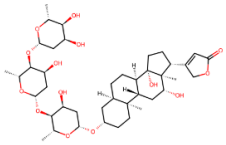

FLUTICASONE PROPIONATE
CAS: 80474-14-2

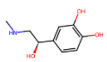

DEXAMETHASONE PHOSPHATE SODIQUE EXPRIME EN DEXAMETHASONE PHOSPHATE
CAS: 2392-39-4

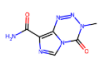

ATROPINE
CAS: 5908-99-6

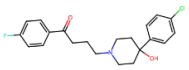

ROSUVASTATINE CALCIQUE
CAS: 147098-20-2

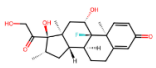

CLOBETASOL PROPIONATE
CAS: 25122-46-7

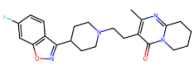

DIGOXINE
CAS: 20830-75-5

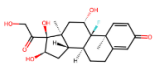

ADRENALINE
CAS: 51-43-4

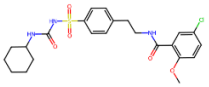

TEMOZOLOMIDE
CAS: 85622-93-1

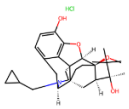

HALOPERIDOL
CAS: 52-86-8

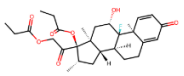

BETAMETHASONE
CAS: 378-44-9

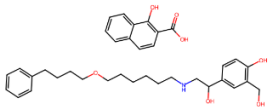

RISPERIDONE
CAS: 106266-06-2

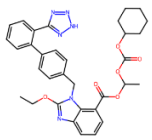

TRIAMCINOLONE / NEOMYCINE
CAS: 124-94-7

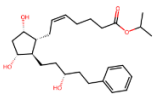

GLIBENCLAMIDE
CAS: 10238-21-8

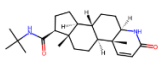

BUPRENORPHINE
CAS: 53152-21-9

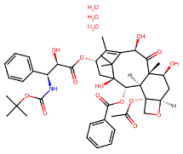

BETAMETHASONE DIPROPIONATE
CAS: 5593-20-4

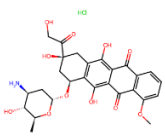

SALM TEROL XINAFOATE EXPRIM EN SALM TEROL
CAS: 94749-08-3

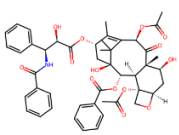

CANDESARTAN CILEXETIL
CAS: 145040-37-5

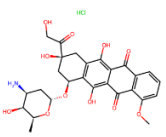

LATANOPROST
CAS: 130209-82-4

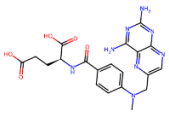

FINASTERIDE
CAS: 98319-26-7


DOCETAXEL TRIHYDRATE
CAS: 148408-66-6


EPIRUBICINE SOUS FORME DE CHLORHYDRATE
CAS: 56390-09-1


PACLITAXEL
CAS: 33069-62-4


DOXORUBICINE
CAS: 23214-92-8


METHOTREXATE
CAS: 59-05-2

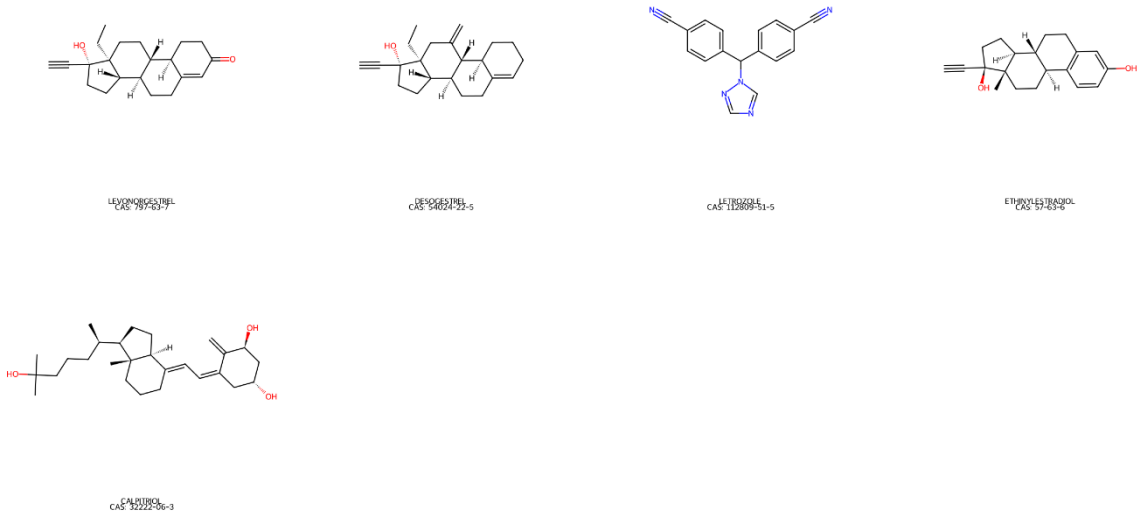


Figure III .4: 221 Molecules Data Representation.

3.2. Feature Extraction

To numerically represent each molecule, the SMILES notation of each API was used to compute two types of molecular features:

- **Molecular descriptors**: A set of 2D physicochemical and topological descriptors computed using **RDKit**.
- **Molecular fingerprints**: 1024-bit **Morgan fingerprints** (ECFP4), which capture substructural patterns within the molecule.

These features were concatenated and reshaped into a 2D array of shape (32×32×2), where one channel corresponds to normalized descriptors and the other to binary fingerprints. This representation was used as input for a **convolutional neural network (CNN)** trained to extract high-level features from chemical structure.

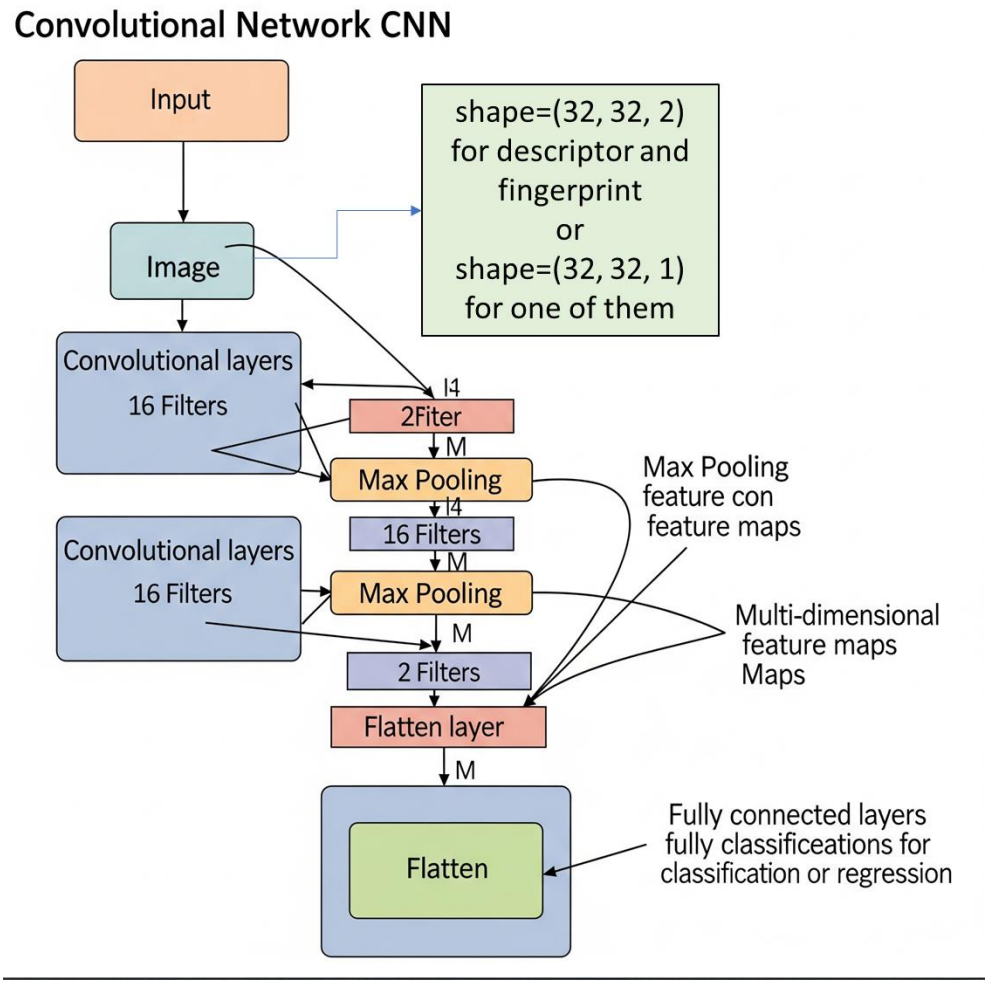


Figure III .5: The diagram shows the structure of the CNN, including the input layer, convolutional layers, max pooling layers, and the flatten layer.

3.3. Model Architecture

The prediction task was formulated as a **multi-class classification problem**. The CNN acted as a **feature extractor**, and its output was passed to classical machine learning classifiers. Five distinct classifiers were evaluated:

- **Multilayer Perceptron (MLP)**
- **Support Vector Classifier (SVC)**
- **Random Forest Classifier**
- **XGBoost Classifier**
- **Decision Tree Classifier**

These classifiers were trained on the CNN-extracted feature vectors and tasked with predicting the correct OEB class for each compound.

4.Hyperparameter Optimization with Optuna

To ensure optimal performance of each classification algorithm, hyperparameter tuning was carried out using Optuna, a modern and efficient hyperparameter optimization framework based on Bayesian optimization. Optuna operates by iteratively sampling parameter values, evaluating model performance, and updating its search strategy based on previous trials, thereby balancing exploration and exploitation efficiently.

For each classifier, a tailored objective function was defined to guide the optimization process. The function receives a trial object and returns a model performance score (in this case, classification accuracy) on the test set. The optimization was conducted over 30 trials per model, and the best-performing parameter set was retained.

The following configurations were explored per model:

MLP Classifier: The number of hidden layers (1–3) and the number of neurons per layer (32–256) were dynamically suggested per trial.

SVC: The regularization parameter C (0.1–100) and kernel coefficient gamma (0.1–0.8) were optimized using the RBF kernel.

XGBoost: Parameters tuned included the number of estimators (50–200), tree depth (3–10), and learning rate (0.01–0.3), with evaluation based on multiclass log-loss.

Random Forest: The number of trees (50–200) and maximum depth (3–15) were varied.

Decision Tree: The maximum tree depth (3–15) and minimum number of samples required to split a node (2–10) were explored.

After identifying the best hyperparameters for each model, the final classifier was retrained on the full training set and persisted for future use. Performance metrics such as accuracy, classification report, and confusion matrix were computed and visualized to assess predictive power.

This approach ensured a standardized and reproducible optimization strategy across all models while leveraging Optuna’s flexibility and computational efficiency for automated model selection.

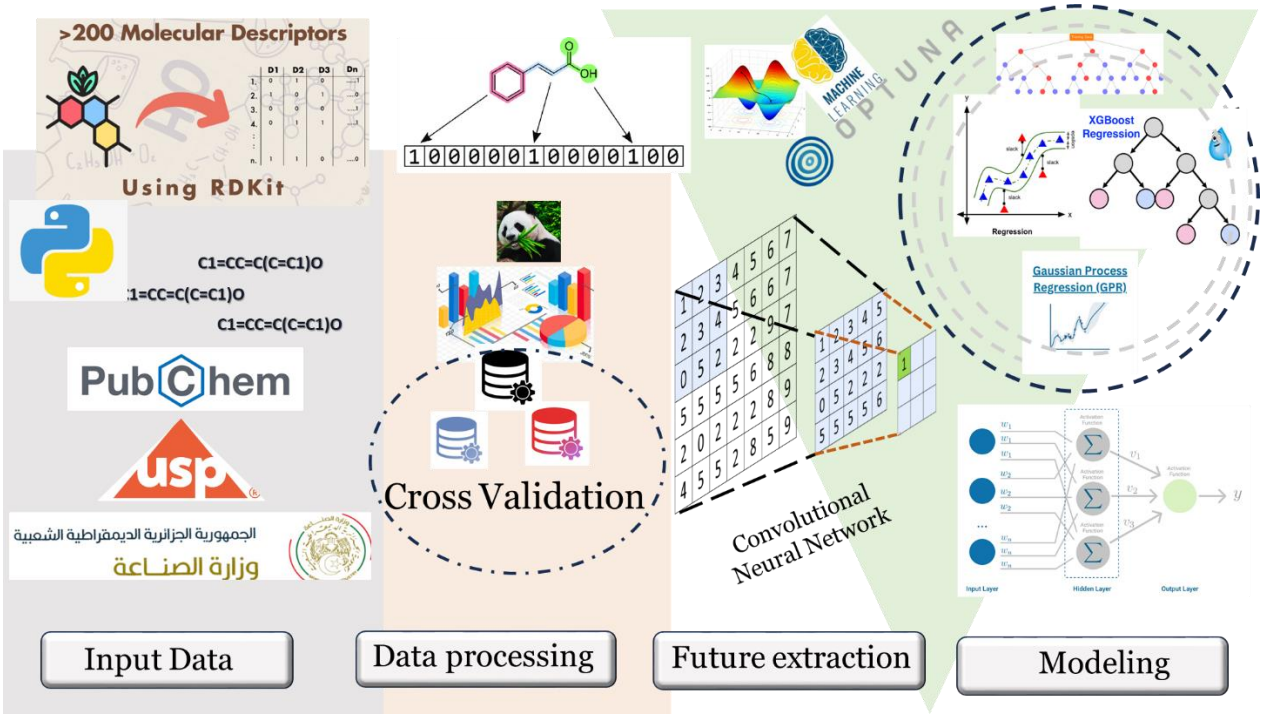


Figure III .6: Modeling process.

5.Evaluation Metrics

To rigorously assess model performance in predicting Occupational Exposure Bands (OEBs), several evaluation metrics were employed: accuracy, precision, recall, and F1-score, along with their macro and weighted averages. These metrics were chosen to provide both a global and class-specific understanding of classification quality, particularly in the presence of class imbalance.

- **Precision** measures the proportion of true positive predictions among all predicted positives for a given class. High precision indicates that the model avoids false positives.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

- **Recall** (also called sensitivity) is the proportion of actual positives that were correctly identified by the model. High recall indicates that the model detects most instances of a given class.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

- **The F1-score** is the harmonic mean of precision and recall, providing a single value that balances both. It is especially useful when the class distribution is imbalanced.

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- **Support** refers to the number of actual instances for each class in the dataset. It indicates how well the model performs relative to the number of samples in each OEB category.

Additionally, the macro average calculates the unweighted mean of a metric across all classes, treating each class equally, whereas the weighted average takes into account the support (sample size) of each class, offering a performance score that reflects class imbalance. Finally, accuracy is the overall proportion of correct predictions, though it may be biased toward majority classes in unbalanced datasets. Collectively, these metrics offer a robust framework for evaluating classifier performance across all OEB categories, especially critical when dealing with rare but hazardous classes like OEB 6.

- Macro average is the unweighted mean of a given metric (e.g., precision, recall, or F1-score) computed independently for each class.

$$\text{Macro Avg} = \frac{1}{N} \sum_{i=1}^{N} \text{Metric for class } i$$

- Accuracy is the proportion of correct predictions (both true positives and true negatives) out of all predictions.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

- The weighted average is the mean of the metric weighted by the number of instances in each class (i.e., class "support").

$$\text{Weighted Avg} = \frac{\sum_{i=1}^{N} (\text{Support}_i \times \text{Metric}_i)}{\sum_{i=1}^{N} \text{Support}_i}$$

In our case:

- Macro avg helps you see whether your model performs fairly across all OEB classes, even the rare ones.
- Weighted average tells you how good the model is when taking class imbalance into account (i.e., it's closer to accuracy but includes precision/recall/F1).
- Accuracy is your overall hit rate, but it may hide poor performance on small classes like OEB 6.

6.Results and Discussion

This section presents the performance of each classification model individually, with a focus on accuracy, class-wise metrics, and confusion matrix analysis. All models were trained using CNN-extracted features from a combined vector of RDKit descriptors and Morgan fingerprints. The evaluation was done on a hold-out test set of 221 molecules, distributed across six OEB classes.

6.1.CNN-Extracted Features

The violin plot illustrates the distribution of normalized feature values extracted by the CNN model across different OEB classes. Most features fall within a relatively narrow range (-0.05 to 0.30), indicating that the model produces compact, low-magnitude representations. While some features show distinct patterns for certain classes—particularly higher-risk bands (OEB 4–6)—many exhibit significant overlap, especially among lower-risk classes (OEB 0–2). This suggests that the CNN may struggle to differentiate molecules in these lower-risk categories, possibly due to inherent similarities in their molecular properties or insufficient discriminatory power in the extracted features.

Notably, a subset of features (primarily in the middle range of the x-axis) demonstrates clearer separation between classes, highlighting their potential importance for classification. The presence of a few outliers with negative values may warrant further investigation to determine whether they represent rare molecular patterns or preprocessing artifacts.

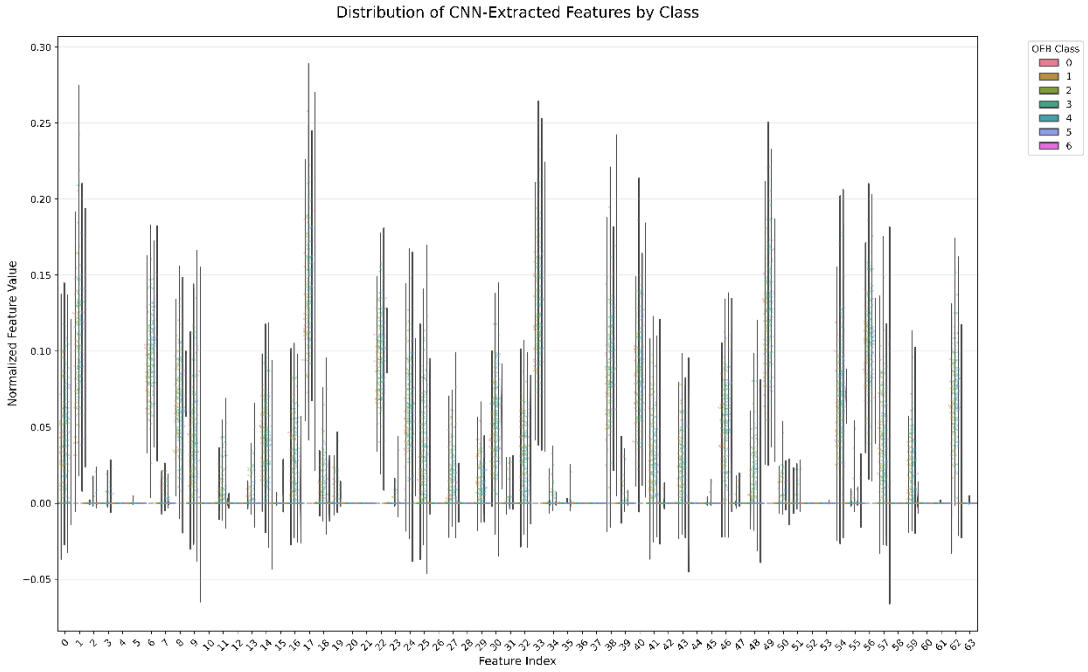


Figure III .7: Distribution of CNN-Extracted Features by OEB Class.

Violin plots show the normalized values of features learned by the CNN model, stratified by OEB classification band. Wider sections indicate a higher density of feature values for a given class. While

some features discriminate higher-risk bands (OEB 4–6), lower-risk bands (OEB 0–2) exhibit substantial overlap, suggesting challenges in distinguishing these classes. Outliers (values < -0.05) may reflect rare molecular patterns or normalization artifacts.

The correlation matrix heatmap provides valuable insights into the relationships between the 64 features extracted by the CNN model. The color gradient represents Pearson correlation coefficients ranging from -1.0 (perfect negative correlation, shown in dark blue) to +1.0 (perfect positive correlation, shown in bright yellow/red). Several important patterns emerge from this visualization that inform our understanding of the model's learned representations.

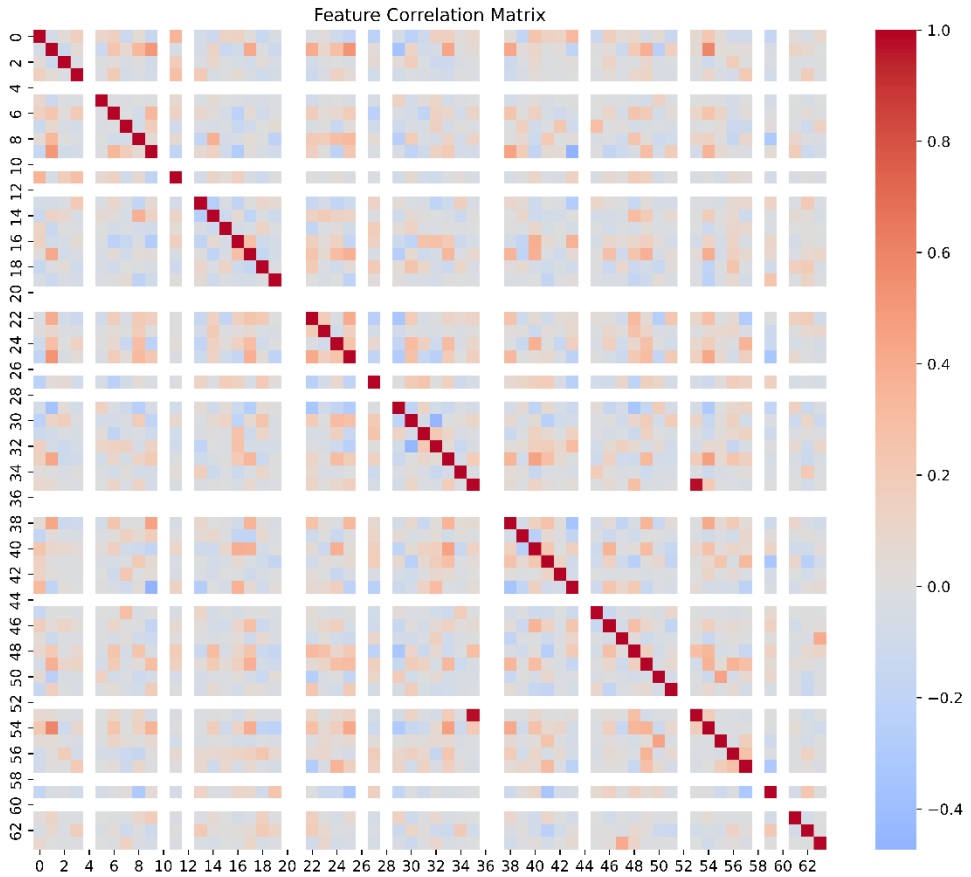


Figure III .8: Correlation Matrix of CNN-Extracted Features.

Heatmap displays Pearson correlation coefficients between 64 learned features. Yellow/red indicates positive correlation (redundant features), blue shows anti-correlation (complementary features), and white/gray represents independent features. Block structures suggest hierarchical feature learning in the CNN architecture.

6.2. Multilayer Perceptron (MLP)

The MLP model achieved an overall **accuracy of 89.6%**. It performed consistently well across most classes, with particularly strong precision for class 1 (100%) and class 6 (100%). However, recall for class 6 dropped to **66.7%**, suggesting under-detection of this minority class.

Table III .2: MLP Macro and Weighted Averages.

MLP_report	***precision***	***recall***	***F1-score***	***Support***
0	0.82	0.90	0.86	10
1	1.00	0.87	0.93	23
2	0.87	0.93	0.90	56
3	0.90	0.90	0.90	67
4	0.88	0.90	0.89	42
5	0.94	0.85	0.89	20
6	1.00	0.67	0.80	3
Accuracy	0.896	0.896	0.896	0.896
Macro avg	0.92	0.86	0.88	221
Weighted avg	0.90	0.90	0.90	221

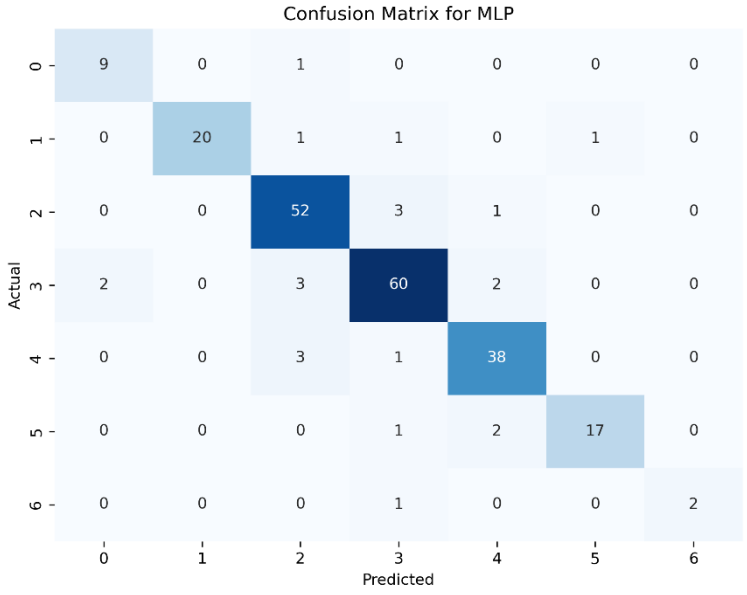


Figure III.9: Confusion Matrix Analyses.

The confusion matrix (**Figure III.9:** Confusion Matrix Analyses.) shows strong diagonal dominance, indicating accurate predictions across classes 0–5. However, class 6 was occasionally misclassified as class 4 or 5, which may be due to low sample size and feature similarity.

Interpretation: MLP demonstrates high generalization for balanced classes but may benefit from data augmentation or penalized learning for rare categories like OEB 6.

6.3. Support Vector Classifier (SVC)

The Support Vector Classifier (SVC) demonstrated strong and balanced performance in predicting OEB classes, achieving an overall **accuracy of 90.5%**. The model maintained consistently high values of **precision (90.7%)**, **recall (90.5%)**, and **F1-score (90.5%)** when averaged across all samples (weighted average). Notably, **the macro-averaged F1-score** was **0.891**, indicating good performance across all classes, including minority ones.

Table III .3: XGBoost report.

	precision	***recall***	***F1-score***	***support***
0	0.900	0.900	0.900	10
1	0.952	0.870	0.909	23
2	0.869	0.946	0.906	56
3	910	910	910	67
4	0.902	0.881	0.892	42
5	0.947	0.900	0.923	20
6	1,000	0.667	0,800	3
accuracy	0.905	0.905	0.905	0.905
macro avg	0.926	0.868	0,891	221
weighted average	0.907	0.905	0.905	221

Class-wise, the SVC performed particularly well on the most populated classes, such as **OEB 2 and 3**, with F1-scores of **0.906** and **0.910**, respectively. Class 5 also saw a high F1-score of **0.923**, reflecting accurate identification of high-hazard compounds. The model was able to **perfectly identify all instances of class 6** (precision = 1.000), although **recall was limited to 66.7%**, likely due to the very small number of samples in this class (n = 3).

The confusion matrix for the SVC model (**Figure III.10**) shows a clear diagonal dominance, confirming accurate classification for most classes. Misclassifications were minimal and primarily occurred between adjacent OEB bands, such as between class 4 and class 5, which may share similar physicochemical profiles. The limited recall for class 6 suggests that while the model is cautious in predicting this rare class (high precision), it may benefit from additional data or oversampling techniques to improve sensitivity.

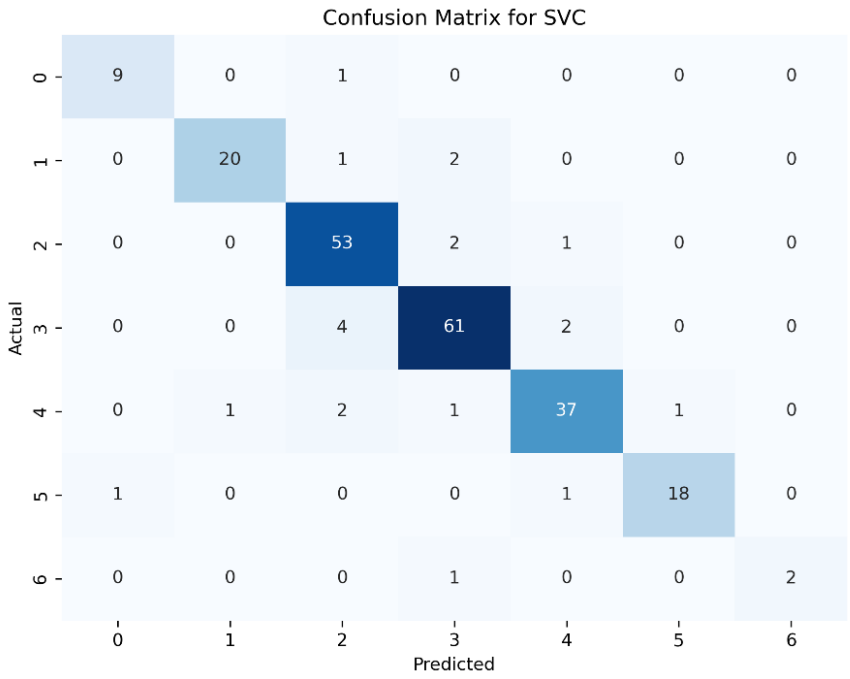


Figure III.10: SVC Confusion Matrix Analysis.

6.4. XGBoost Classifier

XGBoost outperformed all other models with an accuracy of 91.4% and the highest macro-average F1-score (0.905). It combined high recall (94%) for class 3 and perfect precision (100%) for classes 0, 1, and 6.

Table III .4: Support Vector Classifier report.

	precision	*recall*	*F1-score*	*Support*
0	1,000	0.900	0.947	10
1	1,000	0.913	0.955	23
2	0.927	0.911	0.919	56
3	0.863	0,940	0.900	67
4	0.907	0.929	0.918	42
5	0.944	0.850	0.895	20
6	1,000	0.667	0,800	3
accuracy	0.914	0.914	0.914	0.914
macro avg	0.949	0.873	0.905	221
weighted average	0.917	0.914	0.914	221

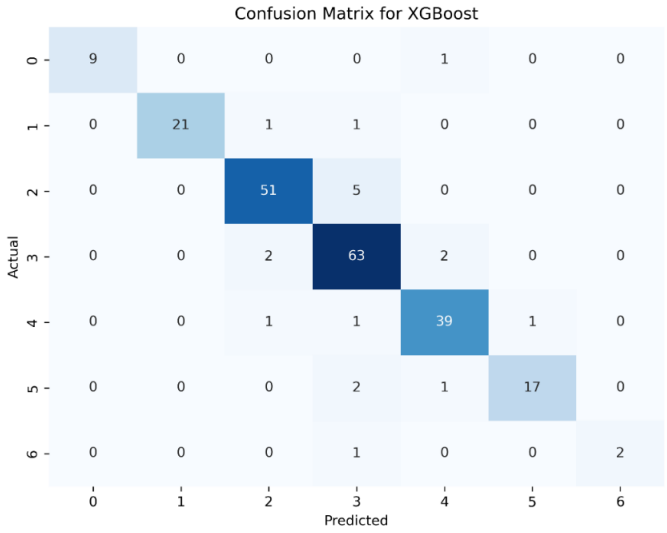


Figure III.11: XGBoost Confusion Matrix Analysis.

The XGBoost confusion matrix (**Figure**) shows excellent class separation and minimal misclassification. Only **a few instances from classes 5 and 6** are slightly misaligned. This suggests that the model effectively learns hierarchical toxicity relationships embedded in the data.

XGBoost offers an optimal trade-off between precision and recall and is especially effective in scenarios where data is imbalanced yet high-dimensional, making it the top-performing model in this study.

6.5. Random Forest Classifier

The **Random Forest Classifier** achieved an overall **accuracy of 90.0%**, indicating strong predictive performance across most OEB classes. It maintained a **macro-average F1-score of 0.896** and a **weighted F1-score of 0.901**, showing that the model performs well not only on frequent classes but also balances minority class predictions with reasonable effectiveness.

Table III .5: Random Forest report.

	Precision	***Recall***	***F1-score***	***Support***
0	1,000	0.900	0.947	10
1	1,000	0.870	0.930	23
2	0.881	0.929	0.904	56
3	0.851	0,940	0.894	67
4	0.900	0.857	0.878	42
5	1,000	0.850	0.919	20
6	1,000	0.667	0,800	3
accuracy	0.900	0.900	0.900	0.900
macro avg	0.948	0.859	0.896	221
weighted average	0.906	0.900	0.901	221

Class-wise performance was particularly high for OEB classes 0, 1, 2, and 5, with precision values of 1.000 and recall values above 85%. The model showed slightly reduced recall for OEB class 4 (85.7%) and the lowest recall for OEB class 6 (66.7%), which again reflects the challenge of predicting underrepresented classes.

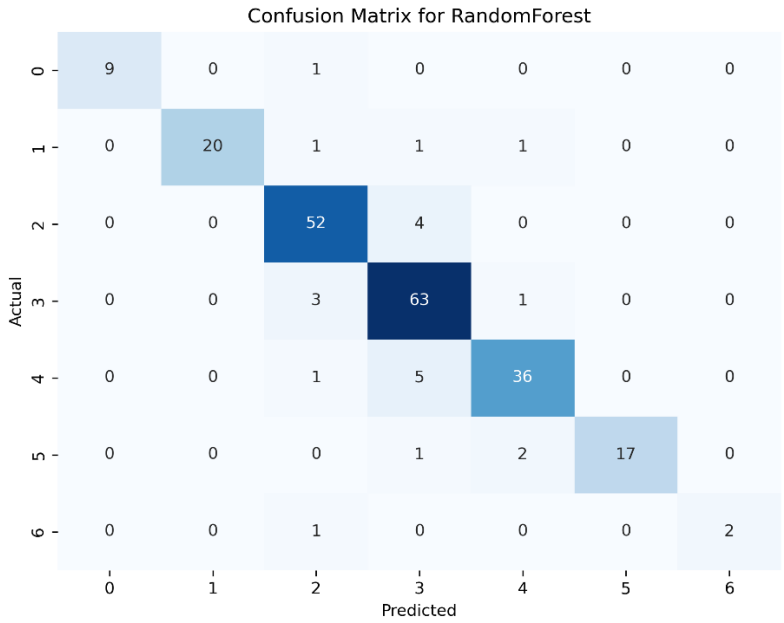


Figure III .12: Random Forest Classifier Confusion Matrix Analysis.

In the confusion matrix for Random Forest (Figure), correct predictions dominate the diagonal, particularly for classes 2 and 3. However, a few misclassifications are observed between **class 4 and class 5**, suggesting structural or toxicological similarities that confuse the classifier. Despite the very limited sample size of class 6, the model still identified two out of three cases correctly.

Random Forest, as an ensemble of decision trees, offers good generalization and resistance to overfitting. Its ability to handle high-dimensional, non-linear data makes it well-suited to toxicological applications. However, performance in rare classes such as OEB 6 may benefit from synthetic sampling or cost-sensitive training.

6.6. Decision Tree Classifier

The decision tree classifier also yielded an accuracy of 90.0%, performing comparably to random forest. Its macro-average F1-score reached 0.893, with a slightly lower weighted F1-score of 0.901. Despite being a simpler, single-tree model, it achieved excellent precision and recall across several classes, especially class 2, where it attained a perfect F1-score (0.929).

Notably, the model showed high performance on classes 0, 1, and 4 and handled class 3 with an F1-score of 0.878. Like other models, it struggled with OEB class 6, predicting two out of three samples correctly (recall = 66.7%, F1 = 0.8).

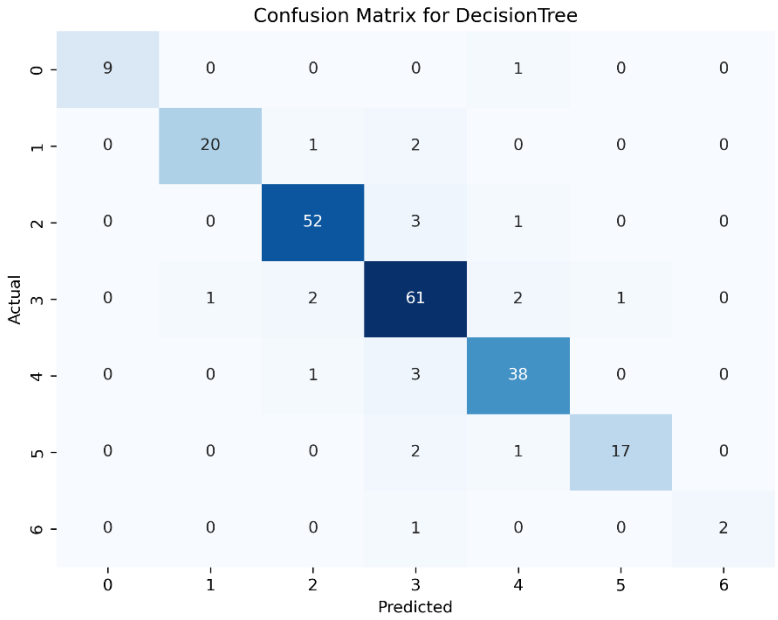


Figure III .13 : Decision Tree Classifier Confusion Matrix Analysis.

The confusion matrix for the decision tree model (Figure 3.13) reflects its strength in cleanly separating most classes. However, like Random Forest, there is slight confusion between adjacent OEB classes, particularly between 3 and 4 and 4 and 5. This suggests overlapping toxicological features that a single decision path may struggle to disentangle.

Table III .6: Decision Tree Report.

	Precision	***Recall***	***F1-score***	***Support***
0	1,000	0.900	0.947	10
1	0.952	0.870	0.909	23
2	0.929	0.929	0.929	56
3	0.847	910	0.878	67
4	0.884	0.905	0.894	42
5	0.944	0.850	0.895	20
6	1,000	0.667	0,800	3
Accuracy	0.900	0.900	0.900	0.900
Macro avg	0.937	0.861	0.893	221
Weighted avg	0.904	0.900	0.901	221

While less complex than ensemble models, the decision tree classifier performs remarkably well given its interpretability and speed. It is particularly useful in contexts where explainability is essential, such as regulatory toxicology. However, its performance on imbalanced classes can be improved through pruning strategies or hybrid modeling.

7.Comparative Analysis of Classification Models

When comparing the performance of all five classifiers (Figure), several key patterns emerge. Overall, the **XGBoost classifier** outperformed the others with the **highest accuracy (91.4%)**, **macro-average F1-score (0.905)**, and strong class-wise performance across all OEB categories, including high-risk bands like OEB 5 and 6. This demonstrates its capacity to effectively handle both class imbalance and feature complexity, making it particularly suitable for toxicological prediction tasks.

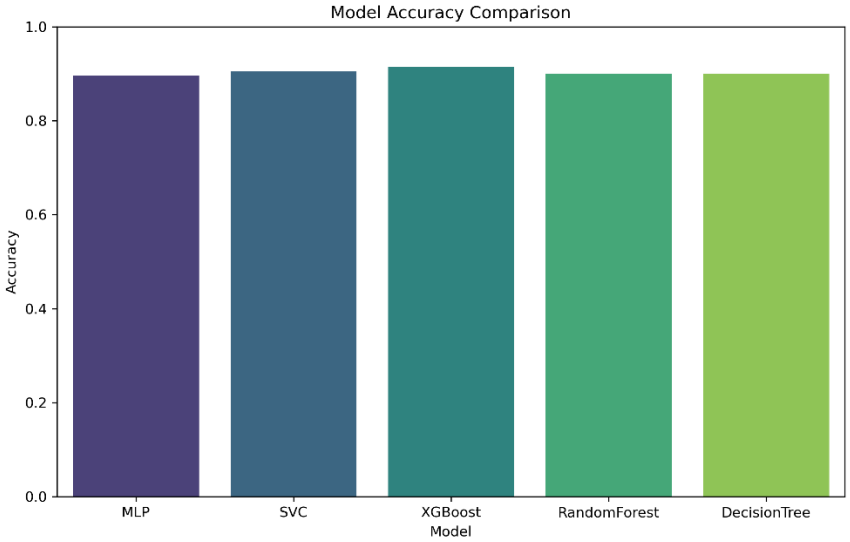


Figure III .14: Comparative Analysis of Classification Models.

The **Support Vector Classifier (SVC)** followed closely with an accuracy of **90.5%**, exhibiting a well-balanced trade-off between precision and recall. It performed consistently across the majority of classes and showed cautious yet correct predictions for the rare OEB 6 class. **Random Forest** and **Decision Tree classifiers** both achieved **90.0% accuracy**, with slight differences in macro F1-scores (0.896 and 0.893, respectively). While Random Forest benefited from ensemble averaging, the decision tree offered nearly equivalent performance with the added benefit of interpretability.

The **Multilayer Perceptron (MLP)** model achieved the lowest accuracy among the models tested, at **89.6%**, though it still performed reliably across most classes. Its limitations were primarily evident in predicting minority classes like OEB 6, where recall was notably lower.

In terms of handling **class imbalance**, XGBoost and SVC showed superior generalization on both majority and minority classes, while MLP and decision trees were slightly less sensitive in correctly identifying rare classes. Despite minor variations, all models demonstrated strong potential in classifying OEB bands from molecular features, confirming the effectiveness of combining **RDKit descriptors**, **Morgan fingerprints**, and **CNN-based feature extraction** in AI-driven toxicological modeling.

8.Graphical User Interface (GUI)

To facilitate user-friendly interaction with the predictive models, a web-based Graphical User Interface (GUI). The application, titled "OEB Prediction Pro," enables users to estimate the Occupational Exposure Band (OEB) of pharmaceutical compounds based on their molecular structure. The GUI is cleanly divided into interactive sections, allowing users to either search for a compound by name via PubChem API integration or directly input a SMILES string to represent the chemical structure.

The sidebar provides a model selection menu, enabling users to choose between five trained machine learning classifiers (MLP, SVC, XGBoost, Random Forest, and Decision Tree). Additionally, a reference table offers quick insight into the definitions of OEB classes, improving interpretability for non-expert users.

Once a molecule is submitted, the application performs real-time feature extraction using RDKit to compute molecular descriptors and fingerprints, followed by transformation through a Convolutional Neural Network (CNN) feature extractor. The resulting features are then passed to the selected classifier, and the predicted OEB class is displayed along with a probability distribution table for all possible classes. A dynamic molecule viewer renders the 2D structure of the input compound, providing immediate chemical context.

The GUI also includes robust error handling for invalid SMILES or missing models and allows visualization of model predictions through styled tables and bar plots. By integrating cheminformatics, deep learning, and intuitive interface design, the GUI serves as a powerful decision-support tool for researchers and regulatory professionals assessing exposure risk in pharmaceutical development.

The interface allows users to predict the Occupational Exposure Band (OEB) of a pharmaceutical compound using a SMILES string or compound name. The user can select from multiple machine learning models, visualize the molecular structure, and view the predicted class along with the probability distribution across all OEB categories. The left sidebar provides model selection and OEB reference information, while the main panels support compound input and result visualization.

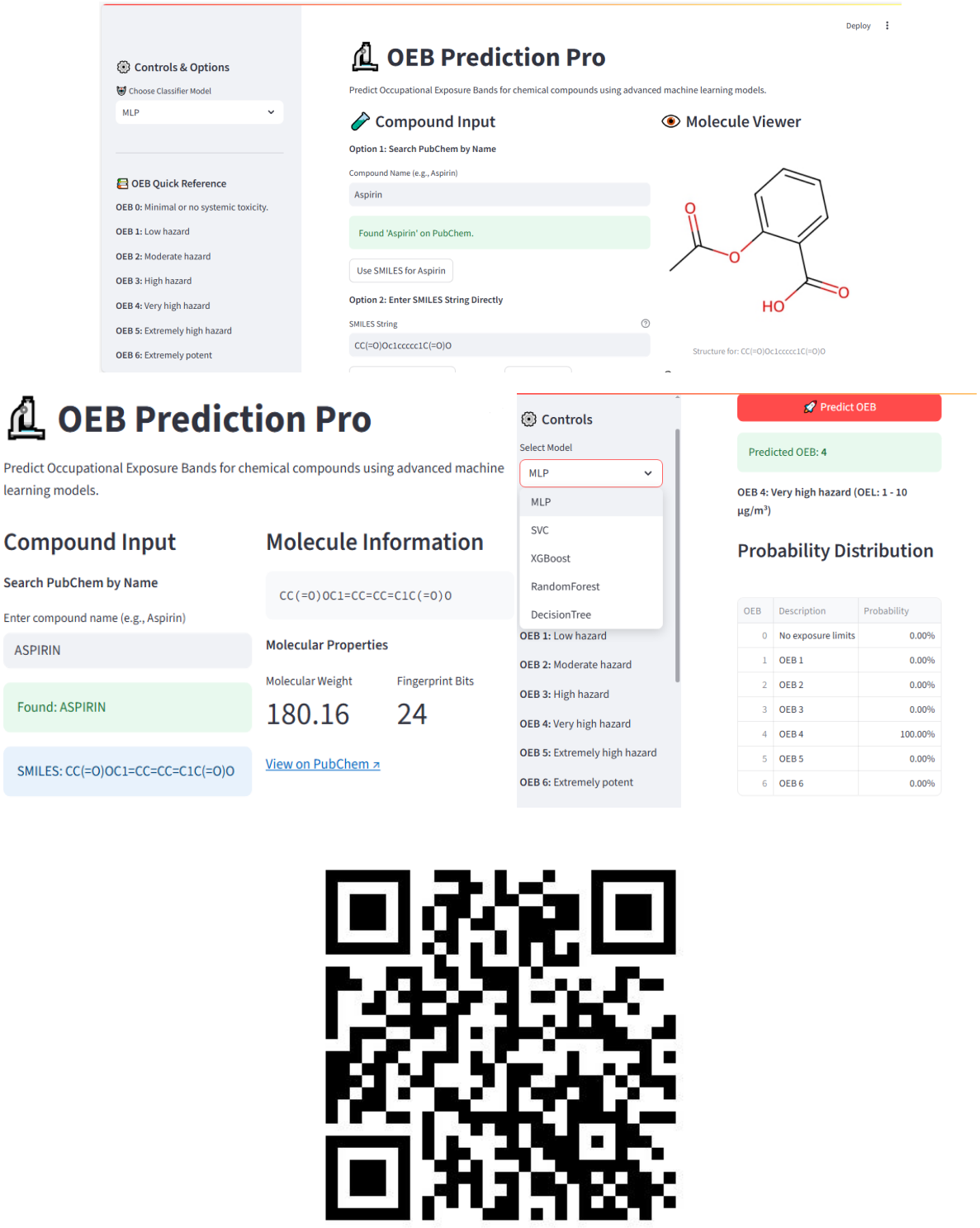


Figure III .15: Screenshot of the OEB Prediction Pro application GUI. Scan the QR code to use online.

8.Conclusion

This chapter presented a comprehensive artificial intelligence-based framework for the prediction of Occupational Exposure Bands (OEBs) for pharmaceutical compounds using cheminformatics techniques. By combining structural representations from SMILES with RDKit descriptors and Morgan fingerprints, a dual-channel feature matrix was constructed to represent each molecule. A CNN architecture was employed to extract deep, informative features, which were subsequently classified using multiple machine learning algorithms.

Among the evaluated models, XGBoost emerged as the top-performing classifier, achieving a 91.4% accuracy and demonstrating robust handling of both majority and minority OEB classes. SVC and Random Forest also exhibited strong performance, closely following XGBoost, while MLP and Decision Tree showed competitive but slightly lower scores. In particular, all models struggled with the highly imbalanced OEB class 6, highlighting the need for further work on data balancing and rare-class sensitivity.

To enhance accessibility and usability, the chapter also introduced a Graphical User Interface (GUI) application—OEB Prediction Pro—that allows users to input compounds by name or SMILES, visualize molecular structures, and obtain real-time OEB predictions with probability distributions. The platform integrates advanced cheminformatics, CNN-based deep learning, and interactive visualization tools, providing a powerful and intuitive decision-support system for pharmaceutical safety assessment.

Conclusion

CONCLUSION

Ensuring safe working conditions in pharmaceutical environments has become increasingly complex due to the rise of highly potent therapeutic compounds. These agents, effective at low doses, can still pose substantial health threats to exposed individuals during production, handling, or administration. The accurate estimation of Occupational Exposure Limits (OELs) is therefore essential for preventing occupational diseases and maintaining workplace safety—especially when empirical toxicity data is lacking.

This research addresses that need by developing a robust AI-based framework for estimating Occupational Exposure Bands (OEBs) using molecular-level data. By integrating cheminformatics tools with deep learning techniques, the study constructed a dual-channel feature representation from molecular descriptors and fingerprints. A Convolutional Neural Network (CNN) was used to extract complex structural patterns, which were subsequently analyzed using various classification algorithms.

Among these, XGBoost achieved the highest performance, with 91.4% accuracy and strong predictive capability across different exposure bands. SVC and Random Forest models also performed well, while MLP and Decision Trees were slightly less consistent. A notable limitation across all models was the difficulty in identifying the rarest and most hazardous band (OEB 6), highlighting ongoing challenges in class-imbalanced datasets.

To translate this research into practical application, a user-friendly software tool—**OEB Prediction Pro**—was developed. It allows users to input molecular data, view compound structures, and receive real-time exposure band predictions along with confidence scores, making advanced toxicological assessment more accessible and actionable.

Overall, this study provides a meaningful step forward in modernizing occupational risk assessment using AI. It supports early safety evaluation in the absence of traditional toxicological data, promotes informed decision-making, and contributes to the broader goal of preventing workplace-related health risks in the pharmaceutical sector.

References

[1]: Martini, M. C. (1992). Landrigan P, Baker D: The recognition and control of occupational disease. JAMA 266:676–680, 1991.: American Journal of Contact Dermatitis, 3(1), 49. https://doi.org/10.1097/01634989-199203000-00019

[2] : Pourbabaki, R., Amin Rajizadeh, M., Faghihi Zarandi, A., Sadeghi-Yarandi, M., & Damiri, Z. (2025). Chemical agents that cause occupational diseases: Toxicity, exposure routes, and health effects. In H. Gül (Éd.), Public Health (Vol. 3). IntechOpen. https://doi.org/10.5772/intechopen.1008774

[3]: United Nations. (2007). *Globally harmonized system of classification and labelling of chemicals (GHS)* (2nd ed.). https://www.unece.org/fileadmin/DAM/trans/danger/publi/ghs/ghs_rev02/English/ST-SG-AC10-30-Rev2e.pdf

[4]: Government of Canada, C. C. for O. H. and S. (2025, juin 2). Ccohs : Whmis - hazard classes and categories. https://www.ccohs.ca/oshanswers/chemicals/whmis_ghs/hazard_classes.html

[5]: Martini, M. C. (1992). Landrigan P, Baker D : The recognition and control of occupational disease. JAMA 266:676–680, 1991.: American Journal of Contact Dermatitis, 3(1), 49. https://doi.org/10.1097/01634989-199203000-00019

[6]: Cherry, N. (1999). Recent advances : Occupational disease. BMJ, 318(7195), 1397-1399. https://doi.org/10.1136/bmj.318.7195.1397

[7]: Paustenbach, D. J. (2001). The history and biological basis of occupational exposure limits for chemical agents. In R. Harris (Éd.), Patty's Industrial Hygiene (1re éd.). Wiley. https://doi.org/10.1002/0471435139.hyg041

[8]: Schenk, L., Hansson, S. O., Rudén, C., & Gilek, M. (2008). Occupational exposure limits : A comparative study. Regulatory Toxicology and Pharmacology, 50(2), 261-270. https://doi.org/10.1016/j.yrtph.2007.12.004

[9]: Musu, T. (2018). Occupational exposure limits: Uses and limitations in worker protection. In T. Musu & L. Vogel (Eds.), Cancer and work: Understanding occupational cancers and taking action to eliminate them (pp. – [include specific pages]). Brussels, Belgium: ETUI.

[10]: American Conference of Governmental Industrial Hygienists. (2025). Introduction to the Chemical Substances TLVs® [PDF]. In TLVs® and BEIs® Documentation. Retrieved from https://www.acgih.org/science/tlv-bei-guidelines/tlv-chemical-substances-introduction/

[11]: Ashford, N. A., & Caldart, C. C. (2025). Environmental protection laws. In International Encyclopedia of Public Health (p. 37-50). Elsevier. https://doi.org/10.1016/B978-0-323-99967-0.00070-3

[12]: Ku, R. H. (2000). An overview of setting occupational exposure limits (Oels) for pharmaceuticals : Setting appropriate occupational exposure limits is an integral component in assuring the health and safety of workers. Chemical Health & Safety, 7(1), 34-37. https://doi.org/10.1016/S1074-9098(99)00070-2

[13]: Ronald, J. W. (2012). Understanding a safety data sheet (Sds) in regards to process safety. *Procedia Engineering*, *45*, 857-867. https://doi.org/10.1016/j.proeng.2012.08.250

[14]: Greenberg, M. I., Cone, D. C., & Roberts, J. R. (1996). Material safety data sheet : A useful resource for the emergency physician. *Annals of Emergency Medicine*, *27*(3), 347-352. https://doi.org/10.1016/S0196-0644(96)70272-X

[15]: Sharma, A., Singh, A., Verma, A., Malviya, R., & Padarthi, P. K. A. (2023). Potential of AI in the advancement of the pharmaceutical industry. In R. Malviya, S. Sundram, & S. Fuloria (Eds.), Pharmaceutical Industry 4.0: Future, Challenges & Application (pp. 107–144). River Publishers

[16]: Sedkaoui, S. (2018). Data analytics and big data. ISTE.

[17]: Singh, B. N., Roy, A., & Maiti, D. K. (Éds.). (2020). Recent advances in theoretical, applied, computational and experimental mechanics : Proceedings of ictacem 2017. Springer Singapore. https://doi.org/10.1007/978-981-15-1189-9

[18]: **IBM.** (n.d.). *What are classification models?* IBM. Retrieved June 18, 2025, from https://www.ibm.com/topics/classification-models

[19]: *Regression in machine learning*. (2017, décembre 1). GeeksforGeeks. https://www.geeksforgeeks.org/machine-learning/regression-in-machine-learning/

[20]: Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. Statistics and Computing, 14(3), 199-222. https://doi.org/10.1023/B:STCO.0000035301.49549.88

[21]: Schölkopf, B. (avec Smola, A. J.). (2002). *Learning with kernels : Support vector machines, regularization, optimization, and beyond*. MIT Press.

[22]: Murphy, K. P. (2012). Machine learning: A probabilistic perspective. MIT Press.

[23]: Sutton, R. S., & Barto, A. (2020). Reinforcement learning : An introduction (Second edition). The MIT Press.

[24]: Zhu, X. (2005). Semi-supervised learning literature survey. Computer Science, University of Wisconsin-Madison.

https://pages.cs.wisc.edu/~jerryzhu/pub/ssl_survey.pdf

[25]: Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255–260. https://doi.org/10.1126/science.aaa8415

[26]: LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444. https://doi.org/10.1038/nature14539

[27]: Sarker, I. H. (2022). Ai-based modeling : Techniques, applications and research issues towards automation, intelligent and smart systems. SN Computer Science, 3(2), 158. https://doi.org/10.1007/s42979-022-01043-x

[28] : Haykin, S. S. (1999). Neural networks : A comprehensive foundation (2nd ed.). PHI Private Limited.

[29]: Wang, H., Chen, B., Lin, C., & Sun, Y. (2018). Observer-based neural adaptive control for a class of MIMO delayed nonlinear systems with input nonlinearities. *Neurocomputing*, *275*, 1988-1997. https://doi.org/10.1016/j.neucom.2017.10.045

[30]: Karpathy, A. (2015). Generating text with recurrent neural networks. Retrieved from http://karpathy.github.io/2015/05/21/rnn-effectiveness/

[31]: article Training Feed-Forward Neural Networks Using Firefly Algorithm.

[32]: Trenn, S. (2008). Multilayer perceptrons : Approximation order and necessary number of hidden units. *IEEE Transactions on Neural Networks*, *19*(5), 836-844. https://doi.org/10.1109/TNN.2007.912306

[33]: Larose, D. T. (2005). *Discovering knowledge in data: An introduction to data mining* (1st ed). Wiley.

[34]: Buscema, M. (1998). Back propagation neural networks. *Substance Use & Misuse*, *33*(2), 233-270. https://doi.org/10.3109/10826089809115863

[35]: Contributed by David S. Wishart Current Protocols in Bioinformatics (2007) 14.1.1-14.1.9 Copyright C 2007 by John Wiley & Sons, Inc. Introduction to Cheminformatics

[36]: Wegner, J. K., Sterling, A., Guha, R., Bender, A., Faulon, J.-L., Hastings, J., O'Boyle, N., Overington, J., Van Vlijmen, H., & Willighagen, E. (2012). Cheminformatics: The computer science of chemical discovery: Appendix. Communications of the ACM. Retrieved from ACM Digital Library

[37]: Schulzinne, H. (2012). The realities of home broadband: Technical perspective. *Communications of the ACM*, *55*(11), 99-99. https://doi.org/10.1145/2366316.2366336

[38]: Molecular Descriptors in Chemoinformatics, Computational Combinatorial Chemistry, and Virtual Screening

Author(s): Ling Xue and Jurgen Bajorath

[39]: Gozalbes, R., & Pineda-Lucena, A. (2011). Small molecule databases and chemical descriptors useful in chemoinformatics: An overview. *Combinatorial Chemistry & High Throughput Screening, 14*(6), 548–558. https://doi.org/10.2174/138620711795767857

[40]: Tropsha, A. (2010). Best practices for qsar model development, validation, and exploitation. *Molecular Informatics*, *29*(6-7), 476-488. https://doi.org/10.1002/minf.201000061

[41]: Technical report: the NIOSH occupational exposure banding process for chemical risk management., 2019. https://doi.org/10.26616/NIOSHPUB2019132

[42]: Graham, J. C., Hillegass, J., & Schulze, G. (2020). Considerations for setting occupational exposure limits for novel pharmaceutical modalities. *Regulatory Toxicology and Pharmacology*, *118*, 104813.