

الجمهورية الجزائرية الديمقراطية الشعبية

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET
POPULAIRE

وزارة التعليم العالي والبحث العلمي

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA
RECHERCHE SCIENTIFIQUE

جامعة صالح بونبندر قسنطينة 3

UNIVERSITE SALAH BOUBNIDER CONSTANTINE 3

كلية الطب - قسم الصيدلة

FACULTE DE MEDECINE - DEPARTEMENT DE PHARMACIE



**Mémoire de fin d'études en vue de l'obtention du
diplôme de
DOCTEUR EN PHARMACIE**

**Application de l'Intelligence Artificielle à la Prédiction des
Bandes d'Exposition Professionnelle (OEB) dans
l'Industrie Pharmaceutique**

Encadré par :

Pr. LALAOUNA Abd El Djalil

Présenté par :

Chergui Achraf Amir

Nasri Hani Abderrahmane

ANNEE UNIVERSITAIRE : 2024 – 2025

ABSTRACT

Introduction: The pharmaceutical industry's rapid development of Highly Potent Active Pharmaceutical Ingredients (HPAPIs) presents significant occupational safety risks. Timely and accurate hazard assessment is crucial, but traditional methods are often hampered by a lack of toxicological data for novel compounds. The current practice of Occupational Exposure Banding (OEB) is a vital risk categorization tool but can be slow and resource-intensive, creating a bottleneck that struggles to keep pace with drug discovery. There is a need for faster, data-driven approaches to support and streamline OEB assignment.

Method: This study proposes and validates a novel approach using Artificial Neural Networks (ANNs) to predict the OEB class of pharmaceutical compounds from their molecular structure. A dataset of 314 compounds with validated OEB classifications was compiled from Safety Data Sheets. Over 2,000 molecular descriptors were calculated and subsequently optimized to a subset of 334 using a Genetic Algorithm to maximize predictive power. Multilayer Perceptron (MLP) models were then trained and validated for each of the five OEB classes as separate binary classification tasks.

Results: The predictive models demonstrated high accuracy and robustness. On the independent validation sets, the ensemble models achieved F1-scores of 83.33% for OEB1, 85.71% for OEB2, 85.71% for OEB3, and 88.89% for OEB4. The model for OEB5, representing the highest hazard level, achieved a perfect F1-score of 100%, indicating flawless identification of high-risk compounds. The Receiver Operating Characteristic-Area Under the Curve (ROC-AUC) values were consistently high, notably reaching 0.90 for OEB1 and 1.00 for OEB5, confirming excellent discriminatory power.

Conclusion: This research confirms the viability of using ANNs as a rapid and reliable tool for OEB prediction. The proposed model provides an objective, data-driven method that can effectively support and accelerate occupational hazard assessment, enabling more efficient implementation of safety measures in the pharmaceutical industry.

Keywords: Artificial Neural Networks, Occupational Exposure Banding (OEB), Pharmaceutical Safety, Predictive Modeling, Molecular Descriptors, Genetic Algorithm

RÉSUMÉ

Introduction : Le développement rapide des Ingrédients Pharmaceutiques Actifs Hautement Puissants (HPAPIs) dans l'industrie pharmaceutique présente des risques importants en matière de sécurité au travail. Une évaluation des dangers rapide et précise est essentielle, mais les méthodes traditionnelles sont souvent limitées par l'absence de données toxicologiques pour les nouveaux composés. La pratique actuelle du Occupational Exposure Banding (OEB) constitue un outil essentiel de catégorisation des risques, mais elle peut s'avérer lente et consommatrice de ressources, créant ainsi un goulot d'étranglement difficile à adapter au rythme de la découverte de médicaments. Il existe donc un besoin urgent d'approches plus rapides, fondées sur les données, pour faciliter et accélérer l'attribution des classes OEB.

Méthode : Cette étude propose et valide une approche innovante utilisant les Réseaux de Neurones Artificiels (RNA) pour prédire la classe OEB de composés pharmaceutiques à partir de leur structure moléculaire. Un jeu de données de 314 composés avec des classifications OEB validées a été compilé à partir de Fiches de Données de Sécurité. Plus de 2 000 descripteurs moléculaires ont été calculés, puis optimisés en un sous-ensemble de 334 grâce à un Algorithme Génétique afin de maximiser la puissance prédictive. Des modèles de Perceptron Multicouche (MLP) ont ensuite été entraînés et validés pour chacune des cinq classes OEB, considérées comme des tâches distinctes de classification binaire.

Résultats : Les modèles prédictifs ont démontré une grande précision et une robustesse élevée. Sur les ensembles de validation indépendants, les modèles ensemblistes ont atteint des scores F1 de 83,33 % pour OEB1, 85,71 % pour OEB2, 85,71 % pour OEB3 et 88,89 % pour OEB4. Le modèle relatif à OEB5, représentant le niveau de danger le plus élevé, a obtenu un score F1 parfait de 100 %, indiquant une identification sans faille des composés à haut risque. Les valeurs de la courbe ROC-AUC étaient systématiquement élevées, atteignant notamment 0,90 pour OEB1 et 1,00 pour OEB5, confirmant un excellent pouvoir discriminant.

Conclusion : Cette recherche confirme la faisabilité de l'utilisation des RNA comme outil rapide et fiable pour la prédiction des classes OEB. Le modèle proposé constitue une méthode objective et fondée sur les données, capable de soutenir et d'accélérer l'évaluation des risques professionnels, tout en permettant une mise en œuvre plus efficace des mesures de sécurité dans l'industrie pharmaceutique.

Mots-clés : Réseaux de Neurones Artificiels, Occupational Exposure Banding (OEB), Sécurité pharmaceutique, Modélisation prédictive, Descripteurs moléculaires, Algorithme génétique

الملخص

المقدمة: يشكل التطور السريع في إنتاج المكونات الدوائية النشطة ذات الفعالية العالية في قطاع الصناعات الدوائية مخاطر مهنية ملحوظة تتطلب إجراء تقييمات سريعة ودقيقة للمخاطر، غير أن الطرق التقليدية لتقييم المخاطر غالباً ما تواجه قيوداً عملية نتيجة افتقارها إلى بيانات سمية كافية عن المركبات الجديدة، ومن ثم تظل عملية تصنيف مستوى التعرض المهني عملية أساسية لكنها بطبيعتها قد تستهلك وقتاً وجهداً كبيرين وتشكل عنق زجاجة أمام وتيرة اكتشاف وتطوير الأدوية؛ لذلك فإن الحاجة ملحة لتبني أساليب أسرع ومبنية على البيانات لدعم وتيسير عملية إسناد فئات التصنيف.

المنهجية: تقترح هذه الدراسة وتحقيق من فعالية نهج مبتكر قائم على الشبكات العصبية الاصطناعية للتنبؤ بفئة التعرض المهني للمركبات الصيدلانية انطلاقاً من خصائصها البنوية والجزئية، حيث جمعت قاعدة بيانات تضم 314 مركباً مصنفاً حسب فئات التعرض استناداً إلى نشرات بيانات السلامة، وحُسب أكثر من 2000 واصف جزئي ثم خُفض هذا العدد إلى 334 واصفاً مختاراً عبر خوارزمية وراثية بهدف تعظيم القدرة التنبؤية، بعد ذلك طُورت نماذج الإدراك متعدد الطبقات وتم تدريب كل نموذج وتعيينته كمسألة تصنيف ثنائي مستقلة لكل واحدة من الفئات الخمس للتعرض.

النتائج: أظهرت النماذج التنبؤية دقة وموثوقية عالية عند تقييمها على مجموعات تحقق مستقلة؛ فقد حققت النماذج المجمعة معدلات F1 بمقدار 83.33% لفئة 1، و85.71% لكل من الفئة 2 والفئة 3، و88.89% للفئة 4، بينما سجل نموذج الفئة 5 — المُمثل لأعلى مستوى من الخطر — درجة F1 كاملة بلغت 100%، ما يدل على تحديد دقيق وخالي من الخطأ للمركبات عالية المخاطر في العينة المختبرة؛ كما كانت قيم المساحة تحت منحنى ROC مرتفعة باستمرار وبلغت على سبيل المثال 0.90 للفئة 1 و1.00 للفئة 5، مما يعكس قدرة تمييز ممتازة للنماذج.

الخاتمة: تبرز هذه الدراسة على أن الشبكات العصبية الاصطناعية تشكل أداة سريعة وموثوقة للتنبؤ بفئات التعرض، وتقدم طريقة موضوعية قائمة على البيانات يمكن أن تسرع وتدعم عملية تقييم المخاطر المهنية، وبالتالي تُسهم في تنفيذ تدابير السلامة بشكل أكثر كفاءة داخل صناعة الأدوية.

الكلمات المفتاحية: الشبكات العصبية الاصطناعية، تصنيف التعرض المهني، السلامة الصيدلانية، النمذجة التنبؤية، الواصفات الجزيئية، الخوارزمية الوراثية.

Table of Contents

Table of Contents	
List of Figures	I
List of tables	III
List of Appendix.....	IV
List of definitions	V
List of Abbreviations.....	IX
GENERAL INTRODUCTION	1
PART I: LITERATURE REVIEW AND THEORETICAL FRAMEWORK.....	2
CHAPTER 1: Pharmaceutical Industry Safety Landscape: Context And Challenges.....	3
1.1 Overview of the Pharmaceutical Industry	4
1.2 Industry Trends Impacting Occupational Safety.....	7
1.2.1 Digital Transformation and Pharma 4.0	7
1.2.2 Increasing Potency of Pharmaceutical Compounds	7
1.2.3 Shift Toward Personalized Medicine	8
1.2.4 Globalization and Supply Chain Complexity.....	8
1.2.5 Environmental Sustainability Initiatives	8
1.3 Occupational Safety in Pharmaceutical Manufacturing.....	9
1.3.1 Unique Hazards in Pharmaceutical Manufacturing	9
1.3.2 Exposure Pathways and Process Risks.....	10
1.3.3 Consequences of Inadequate Safety Measures.....	10
1.3.4 Case Studies and Success Stories.....	11
1.4 Broader Drivers of Occupational Safety Demand.....	13
1.4.1 Regulatory Scrutiny and Compliance Requirements	13
1.4.2 Corporate Social Responsibility and ESG Initiatives	13
1.4.3 Labor Market Dynamics and Workforce Expectations.....	14
1.4.4 Economic and Business Continuity Considerations.....	14
1.5 General Approaches to Occupational Safety in Pharma	16

1.5.1 Hierarchy of Controls.....	16
1.5.2 Risk Assessment and Management	16
1.5.3 Containment Strategies	17
1.5.4 Exposure Monitoring and Surveillance.....	18
1.6 Occupational Safety Challenges Unique to Pharma	19
1.6.1 Rapid Introduction of New Compounds	19
1.6.2 Confidentiality and Data Limitations.....	20
1.6.3 Balancing Innovation with Safety	20
1.6.4 Multi-product Facilities and Cross-contamination Risks.....	20
1.6.5 Integration with GMP Requirements	21
CHAPTER 2: Containment in the Pharmaceutical Industry	22
2.1 Regulatory Framework for Containment	24
2.2 Historical Background of Containment.....	26
2.3 Principles of Containment.....	29
2.4 Exposure Assessment & Control Strategies.....	31
2.4.1 Occupational Exposure Limits	32
2.4.2 Occupational Exposure Banding (OEB)	34
2.4.3 Exposure Metrics and Measurement	39
2.4.4 Comparison of Metrics and Approaches	42
2.5 Containment Technologies & Strategies.....	46
2.6 Containment Validation and Verification	50
2.7 Current Trends & Future Directions	53
CHAPTER 3: Neural networks and their use in the pharmaceutical industry	57
3.1 Introduction to Artificial Intelligence in Pharmaceutical Safety Assessment.....	58
3.2 Network Architectures in Predictive Modeling	58
3.2.1 Multilayer Perceptron (MLP).....	58
3.2.2 Radial Basis Function Networks (RBF).....	60
3.3 Activation Functions	61

3.4 Optimization Algorithms in Neural Networks	62
3.4.1 Stochastic Gradient Descent (SGD)	63
3.4.2 Adam (Adaptive Moment Estimation)	63
3.4.3 L-BFGS (Limited-memory Broyden–Fletcher–Goldfarb–Shanno)	64
3.5 Training Methodology and Backpropagation	65
3.5.1 Backpropagation	65
3.6 Evaluation of ANN Performance	66
3.6.1 Evaluation Metrics	67
3.6.2 Confusion Matrix Analysis	67
3.7 The advantages of ANN in OEB predictions	69
3.8 Advantages and Limitations of Neural Networks in Hazard Classification	69
PART II: METHODOLOGY, MODEL IMPLEMENTATION, AND PREDICTIVE	
RESULTS.....	71
1. Data Gathering	72
1.1 Introduction	72
1.2 Materials and Methods	73
1.2.1 Data Sources	73
1.2.2 Inclusion and Exclusion Criteria	73
1.2.3 Molecular Descriptor Calculation	74
1.3 Results	75
1.4 Discussion	79
1.5 Conclusion.....	80
2. Data Selection & Cleaning	80
2.1 Introduction	80
2.2 Materials and Methods	81
2.3 Results	82
2.4 Discussion	83
2.5 Conclusion.....	84
3. Descriptor Selection Using Genetic Algorithm	85

3.1 Introduction	85
3.2 Methodology	85
3.2.1 Overview of Genetic Algorithms	85
3.2.2 Implementation for Descriptor Optimization.....	87
3.2.3 Optimization Workflow Applied in This Study.....	88
3.3 Results	88
3.4 Discussion	89
3.5 Conclusion.....	89
4. Data Processing and Predictive Model Training.....	89
4.1 Justification for Using Artificial Neural Networks (ANNs)	89
4.2 Data Preparation.....	90
4.3 OEB Class Distribution Analysis.....	90
4.4 Data Splitting.....	90
4.5 Training Strategy.....	90
4.5.1 Model's selection	91
4.5.2 Loss Function and Optimization Algorithm.....	91
4.5.3 Epochs and Batch Size	91
4.5.4 Monitoring Model Performance.....	91
4.5.5 <i>Visualizing Learning Progress</i>	92
4.6 Model Evaluation	92
4.6.1 Evaluation Metrics	92
4.6.2 Confusion Matrix Analysis	93
4.6.3 ROC Curve and AUC Score.....	93
4.7 Post-Training Analysis	93
4.7.1 Weight Distribution Analysis.....	93
5. Results	94
Overall Project Scope and Model Combinations	94
Model Achievements and Criteria Fulfillment.....	94

5.1 Development of a predictive model for OEB1 class determination.....	94
5.1.1 Results and Performance Evaluation.....	101
5.1.2 Confusion Matrix Analysis	103
5.1.3 Roc-curve interpretation.....	106
5.1.4 Variable importance	107
5.2 Development of a predictive model for OEB2 class determination.....	111
5.2.1 Results and Performance Evaluation.....	116
5.2.2 Confusion Matrix Analysis	119
5.2.3 Roc-curve interpretation.....	121
5.2.4 Variable importance	123
5.3 Development of a predictive model for OEB3 class determination.....	127
5.3.1 Results and Performance Evaluation.....	132
5.3.2 Confusion Matrix Analysis	135
5.3.3 Roc-curve interpretation.....	137
5.3.4 Variable importance	140
5.4 Development of a predictive model for OEB4 class determination.....	144
5.4.1 Results and Performance Evaluation.....	149
5.4.2 Confusion Matrix Analysis	151
5.4.3 Roc-curve interpretation.....	153
5.4.4 Variable importance	155
5.5 Development of a predictive model for OEB5 class determination.....	159
5.5.1 Results and Performance Evaluation.....	164
5.5.2 Confusion Matrix Analysis	168
5.5.3 Roc-curve interpretation.....	170
5.5.4 Variable Importance	171
6. General Conclusion	177
References :	
Appendix :	