REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Mohamed Khider – BISKRA
Faculté des sciences et de la technologie
Département du Génie Electrique

N° d'ordre :……...

THESE

En vue de l'obtention du diplôme de **DOCTORAT LMD** en E**lectronique**

Spécialité :

Biométrie et Télésurveillance

Titre

Approche de prédiction par télésurveillance à base de Data Mining

Présentée par :

Samia Noureddine

Soutenu le : 17/03 /2022

Devant le jury composé de :

Présidente :	Pr. Nadjiba Terki, Université de Biskra.
Directeur de thèse :	Pr. Zine Eddine Baarir , Université de Biskra.
Co- Directeur de thèse :	Pr. Abida Toumi, Université de Biskra.
Examinatrice :	Pr. Leila Hayet Mouss, Université de Batna.
Examinatrice :	Pr. Kinza Nadia Mouss, Université de Batna.

Année universitaire 2021/2022

Résumé

Suite à l'évolution technologique notamment l'approche mobile, la recherche scientifique s'est orientée vers l'exploitation de ces progrès pour l'aide à la décision prédictive à distance. Un intérêt majeur des chercheurs a eu un grand impact dans le domaine médical à cause de son influence très positive pour la prise en charge du malade ayant pour but son assistance et la réduction des cas de décès à cause de suivi et du problème de temps d'actiond'urgence. C'est ainsi que la télémédecine est devenue un enjeu de grande importance, cela repose sur la manipulation et l'analyse d'un grand volume de données médicales.

Le but de cette thèse consiste d'une part à exploiter une nouvelle approche d'analyse de données à savoir Symbiotic Organisms Search (SOS) pour le Data Mining pour la classification des données, d'autre part, à proposer des améliorations de cette métaheuristique. Cette amélioration repose sur l'intégration de la vitesse dans SOS comme un nouveau paramètre pour explorer l'espace de recherche efficacement et en évitant la convergence prématurée. Nous développerons aussi une mise en œuvre d'une architecture conceptuelle et pratique pour la télémédecine appliquée pour l'aide à la décision pour la connaissance du type de cancer du sein (bénin ou malin). Cette étude nous a permis d'aboutir à d'excellent résultat et constat en matière de classification de données.

Mot clés : Data Mining, Classification, Télémédecine, Métaheuristique, Aide à la décision, Apprentissage automatique, Prédictions.

Abstract

Following the technological evolution, in particular the mobile approach, scientific research has been oriented towards the exploitation of these advances for remote predictive decision support. A major interest of researchers has had a great impact in the medical field because of its very positive influence for the care of the patient aimed at its assistance and the reduction of cases of death due to follow-up and the problem of time of treatment. emergency action. This is how telemedicine has become an issue of great importance, it is based on the manipulation and analysis of a large volume of medical data.

The aim of this thesis is firstly to exploit a new approach to data analysis, namely Symbiotic Organisms Search (SOS) for Data Mining for data classification, and secondly, to propose improvements to this metaheuristic. This improvement relies on the integration of speed in SOS as a new parameter to explore the search space efficiently and avoiding premature convergence. We also develop a conceptual and practical architecture for applied telemedicine for decision support for the knowledge of the type of breast cancer (benign or malignant). This study allowed us to achieve excellent results and findings in terms of data classification.

Keywords: Data Mining, Classification, Telemedicine, Metaheuristics, Decision support, Machine learning, Prediction.

الملخص

بعد التطورات التكنولوجية ، ولا سيما النموذج المتنقل ، ركز البحث العلمي على استغلال أوجه التقدم في دعم القرار التنبئي عن بُعد .كان للأهتمام الكبير من الباحثين أثر كبير في المجال الطبي من حيث دعم القرار التنبئي لما له من تأثير إيجابي للغاية في رعاية المريض بهدف مساعدته و التقليل من نسبة الوفيات بسبب اتباعه والتحكم في مشكلة التدخل في الوقت الأحسن والمناسب. هذه هي الطريقة التي أصبح بها العلاج عن بعد قضية ذات أهمية كبيرة ،.

يتمثل الهدف من هذه الأطروحة من ناحية استغلال نهج جديد لتحليل البيانات ألا و هي خوارزمية البحث عن الكائنات التكافلية (SOS) لتحليل البيانات و تصنيفها، و كذلك، تم اقتراح تحسينات على هذه الخوارزمية (SOS) . يعتمد هذا التحسين على تكامل السرعة في SOS كمعامل جديد لاستكشاف مساحة البحث بكفاءة وتجنب التقارب المبكر أتاحت لنا هذه الدراسة تحقيق نتائج ممتازة من حيث تصنيف البيانات. متبوعة بانجاز و تقييم بنية تصميمية لعملية المتابعة الطبية عن بعد مطبقة على دعم القرار لمعرفة نوع سرطان الثدي (حميد او خبيث).

الكلمات المفتاحية : Metaheuristics، التنقيب في البيانات ، التصنيف ، الطب عن بعد، دعم القرار ، التعلم الآلي،التنبؤ

REMERCIEMENTS

Je tiens d'abord à remercier Allah pour m'avoir donné l'énergie et la volonté pour finaliser ce travail de thèse.

Je tiens à remercier le professeur **Zine Eddine Baarir** directeur de cette thèse pour ses valeureux conseils et orientations qui ont largement contribuer à l'aboutissement de ce travail.

Mes remerciements vont également au professeur **Abida Toumi** la co-directrice de cette thèse en reconnaissances de l'apport très précieux ainsi que le suivi très rigoureux qui m'ont permis l'achèvement de ce travail.

Mes remerciements vont aussi vers les membres de jury pour avoir acceptés de m'honorer avec leur présence et m'accorder de précieux remarques. Mon profond respect au professeur **Nadjiba Terki** enseignante à l'université Mohamed Khider Biskra, de nous avoir fait l'honneur d'accepter d'être la présidente du jury.

Mes sincères remerciements et respects au professeur **Leila Hayet Mouss** et au professeur **Kinza Nadia Mouss** de l'université Mostafa Ben Boulaid Batna 2, de m'avoir fait honneur en acceptant d'être membre de jury.

Je ne peux aussi oublier docteur **Abir Betka** pour sa présence durant la mise en valeur des idées scientifiques ainsi que la mise au point des réflexions nécessaires faisant du travail un résultat très bien présenté.

DEDICACES

Je commence d'abord à dédier ce travail à la mémoire de mon défunt cher père, à ma mère pour sa présence à côté de moi.

Je dédie aussi ce travail à :

- A mon mari pour son aide et sa présence prêt de moi.
- Ma fille **Dr. Nesrine Kazar** pour son encouragement et ses efforts envers moi ;
- Mon fils **Nazim Kazar** qui a fait de ma thèse un des objectifs de sa vie ;
- Mon fils **Nizar Kazar** qui a fait de ma thèse une source de joie.

Je ne peux oublier aussi à dédier ce travail à toute ma famille :

- Mes frères : **Djamel, Mohamed** et **Lotfi** qui mérite toute ma reconnaissance et gratitude pour sa présence permanente près de moi.
- Mes sœurs : **Nadjiba, Nora** en particulier **Meriem** qui est plus qu'une sœur et qui mérite une dédicace particulière pour son soutien moral.
- Mes belles sœurs : **Karima** et **Soulaf**.
- Mes neveux et mes nièces en particulier **Dr. Imene Hassani** pour son attention particulière envers moi ;

A mes amies qui m'ont longuement encouragé : **Saida Lamri, Abida Toumi, Nabila Souidi, Soraya Zehani, Farhi Kamilia, Fatiha Toumi, Nadjiba Terki, Amina Saadallah,** et à toutes mes amies.

Liste des figures

Liste des tableaux

Liste des abréviations

ABC : Artificial Bee Colony /Colonie d'abeilles artificielles.

ANN : Artificial Neural Network / Réseau de NeuronesArtificiels.

AOA : Algorithme d'Optimisation d'Archimède.

ASOS : Adaptive SymbioticOrganismsSearch / Recherche d'organismes symbiotiques Adaptatif.

BBO : Biogeography-Based Optimizer / Optimiseur Basé sur la Biogéographie.

BC : Breast Cancer / Base de données due cancer du sein

BOA : Butterfly OptimizationAlgorithm **/** Algorithme d'Optimisation des Papillons**.**

CA : Cultural Algorithms / Algorithmes culturels.

CDWPSO : Chaotic Dynamic WeightParticleSwarmOptimization / Optimisation de l'essaimde particules de poids dynamique chaotique.

DE : Differential Evolution / Evolution Différentielle

DSS : Decision Support System / Système d'Aide à la Décision.

DT : Decision Tree / Arbre de Décision.

ESOS : Enhance Symbiotic Organisms Search / Améliorer la recherche d'organismes symbiotiques.

GA : GeneticAlgorithm / Algorithme Génétique.

GWO : Grey Wolf Optimisation /L'Optimisation du Loup Gris.

GWOLA : Grey wolfOptimizer-Learning Automata / les automates d'apprentissage basés sur l'optimiseur de loup gris.

GWOTS : Gray Wolf OptimizerTabooSearch / hybride combinant l'optimiseur de loup gris et la recherche Tabou.

HD : HeartDisease / Base de données sur les maladies cardiaques

HWBA : Hybrid Wolf-Bat Algorithm / Algorithme hybride loup-chauve-souris.

JS : Jellyfish Search / Recherche de méduses.

K-NN : K-Nearest Neighbours / K-Voisins les plus Proches.

LD : Liver Disorder / Base de données sur les troubles hépatiques.

LR : Logistique Regression /Régression Logistique.

MC : Matrices de Confusion

MLP : Multi Layer Perceptron / Perceptron MultiCouche.

MLR : Multinomial Logistic Regression /Régression logistique multinomiale.

MSE : Mean Square Error / l'erreur quadratique moyenne.

NB : Naïve Bayes.

OMS : Organisation Mondiale de la Santé.

PBIL : Population-BasedIncremental Learning /Apprentissage progressif basé sur la population.

PID : Base de données de Pima Indian Diabetes.

PSO : Particule Swarm Optimisation /Optimisation pour Essaim de Particules.

SOS : Symbiotic Organisms Search / Recherche d'Organismes Symbiotiques.

SPCE : Smart Primary Care Environment / Environnement de Soins Primaires Intelligent.

SVM : Support Vector Machine / Machine à vecteur de soutien.

TIC : Technologies d'Information et de Communication. / Information and Communication Technology.

VSOS : Velocity Symbiotic Organisms Search / Vélocité Recherche d'Organismes Symbiotiques.

TABLE DES MATIERES

INTRODUCTION GENERALE

INTRODUCTION GENERALE

1. Contexte

De nos jours et suite à l'évolution technologique et les technologies de l'information et de communication associés à l'informatique mobile ainsi que l'intelligence artificielle, il est devenu très fréquent leur exploitation dans des domaines variés pour l'amélioration des services fournis aux acteurs (client, décideurs,..).

A cet effet, depuis plusieurs années dans le domaine médicalde des recherches ont eu lieu pour l'optimisation qualitative d'une part des services médicaux développés envers les malades, d'autre part il s'agit de fournir aux experts médicaux les données nécessaires sous formes adaptées à des besoins bien spécifiques pour une prise de décision de qualité en temps réel et parfois à distance.

Cette prise de décision est fondée sur une analyse des données médicales des patients dont le volume et assez grand. C'est ainsi que dans ce domaine d'analyse de données appelé aussi « Data Analytics » se place notre contexte de recherche. Plus exactement, notre travail de recherche s'inscrit dans le domaine du datamining avec une mise en valeur conceptuelle et pratique dans la médecine.

2. Problématique

La prise de décision médicale prédictive repose sur l'analyse d'un grand volume de données de type bigdata. Les données sont générées par des machines de différent type via des capteurs transmit à une base de données centrale. Les experts médicaux prennent des décisions sur la base d'une connaissance brute qui génèrent parfois des confusions médicales.

C'est ainsi que l'extraction de la connaissance nécessaire en temps réel est devenue un enjeu scientifique et économique de grande importance. Cette forme est définie par le concept de télémédecine qui consiste à exploiter des systèmes d'aide à la décision médicaux à distance.

Dans le domaine médical, la télésurveillance et le datamining sont exploitées pour résoudre certains de ses problèmes afin de fournir de bons soins aux patients, parmi lesquels : les maladies chroniques et infectieuses qui nécessitent une surveillance permanente à distance. La position distante du malade par rapport du service d'urgence medical est critique et pourrait devenir une menace sérieuse.

La télémédecine est considérée comme «axée sur les services» en raison de l'adoption de la technologie d'accès sans fil. Le système de décision prédictif est basé sur l'analyse d'une grande masse de données assurant la génération de modèles et des statistiques permettant à des experts d'agir à distance rapidement et efficacement pour le suivi des patients.

3. Motivation

L'aide à la décision prédictive est devenue d'une part un apport très apprécié et très demandé dans le domaine médical, d'autre part son application via une approche à distance est aussi devenue une exigence et une recommandation des acteurs du secteur médical suite aux résultats aux avis des praticienes médicaux.

Notre recherche est motivée essentiellement par le volume important de données générées dans le domaine médical où la décision doit être la plus optimale possible pour éviter toute complication ultérieure concernant le patient. En revanche, l'extraction des connaissances à partir d'un grand volume de données doit répondre à des critères qui varient parfois d'un expert à l'autre ; cela doit toujours répondre à un besoin très précieux et précis en termes de prise de décision.

4. Objectif et contributions

L'objectif de cette thèse porte sur la mise en pratique ainsi qu'une validation théorique et expérimentale de l'importance du Machine Learning en général et du datamining en particulier en tant que processus clé et de qualité dans la télésurveillance médicale.

C'est ainsi que dans cette thèse nous allons définir une approche prédictive d'analyse de données ou datamining dans le domaine médical. Nos contributions porteront sur :

- L'expérimentation d'une nouvelle métaheuristique dite Recherche d'Organismes Symbiotiques (en Anglais SOS : Symbiotic Organisms Search) comme approche qualitative en matière de classification de données et la prédiction pour le domaine médical ;
- L'amélioration de SOS vers une nouvelle forme appelée Vélocité SOS (en Anglais VSOS : Velocity Symbiotic Organisms Search) visant à améliorer son exploration une optimisation de classification des objets. Après expérimentations, nous obtenons une nette amélioration de SOS en termes de résultats obtenus.
- Expérimentation de VSOS sur des données médicales concernant le cancer de sein, ainsi que sur des données médicales sur le domaine cardiaque.
- La modélisation d'une approche de télémédecine pour un datamining médical via la technologie des réseaux de capteurs en mode de télésurveillance pour l'aide à la décision appliquée sur le cancer de sein.

5. Organisation de la thèse

Notre recherche est rédigée dans cette thèse suivant une progression cohérente :

- Dans le premier chapitre nous avons établit un état de l'art sur le domaine du datamining en insistant sur les concepts clés, les approches utilisées ainsi que certains

- Domaines d'application suivi d'une synthèse des travaux réalisés dans le domaine médical.
- Le second chapitre est dédié à notre première contribution, c'est une étude théorique et expérimentale du nouvel algorithme de clustering basé sur une métaheuristique dite Recherche d'Organismes Symbiotiques (SOS),

- Le troisième chapitre comporte notre deuxième contribution, c'est une optimisation de SOS vers VSOS, les résultats obtenus sont très satisfaisants.
- Dans le quatrième et dernier chapitre qui réprésente la troisième contribution, nous avons développé une mise en œuvre de VSOS pour l'aide à la décision appliquée au cancer du sein.

Nous clôturons notre thèse par une conclusion générale synthétisant nos travaux de recherche développés dans ce cadre, par la suite nous discuterons brièvement des perspectives sous forme de verrous scientifiques qui représentent une suite de nos contributions qui méritent une attention scientifique.

Chapitre I

Etat de l’art sur le Data Mining et domaines d'application

I.1 Introduction

Le Data Mining appelé aussi extraction de connaissance à partir d'une grande masse de données est devenu un enjeu économique et scientifique de grande importance. Cet intérêt vient du rôle primordial que procure le datamining pour la prise de décision. Les enjeux et le besoin du datamining sont même devenus très demandé notamment dans des systèmes de prédiction assurant une prise de décision prématurée pour éviter des situations complexes menant à des fins chaotiques [1].

L'avancé technologique est devenue un refuge qualitatif en matière d'acquisition et de stockage de données nécessistant en consèquence une analyse très appréciable et de valeur certaine assurant une conduite post traitement. L'impact de la « data » et du Bigdata a influencé plusieurs travaux dans le domaine du datamining exploitant des techniques de l'intelligence artificielle, les métaheuristiques et l'analyse de données [1].

Dans ce chapitre nous allons donner un aperçu général sur le principe du datamining. Nous donnerons aussi les grands axes de recherche ainsi que les domaines d'application les plus cités dans la littérature.

Ce chapitre est organisé comme suit : nous commençons d'abord par introduire des notions fondamentales du datamining, puis nous donnerons une description exhaustive des approches existantes et appliquées pour le dataming. Nous dresserons aussi une étude comparative entre les méthodes de datamining plus spécialement une comparaison basée sur les caractéristiques des données. En suite, une synthèse de certains travaux en étroite relation avec le domaine d'application à savoir le domaine médical, ainsi qu'une comparaison qualitative des travaux cités seront effectués.

I.2 Généralité sur le Data Mining

Dans cette section, nous allons décrire les principaux concepts liés au domaine du datamining.

I.2.1. Définition du Data Mining

Le datamining est apparu vers les années soixante. Ces dernières années, le dataming est devenu un refuge scientifique pour les décideurs des entreprises en matière de suivi des profils et des actions commerciales des clients et dans plusieurs autres secteurs comme l'industrie, la santé, etc. Le datamining est une méthodologie de base des technologies du Big Data et des techniques d'analyse de données volumineuses. C'est un sous domaine du domaine appelé « Big Data Analytics » dont les principaux traitements sont l'analyse prédictive et l'exploitation des données. Le datamining greprésente aussi l'analyse de données suivant différentes perspectives et qui génére à partir de ces données iniales des connaissances ou des informations

utiles et importantes en établissant des relations entre les données ou en repérant des patterns [1][2].

Les données extraites à partir d'une masse de données sont ensuite utilisées pour une prise de décision par les decideurs d'entreprises pour augmenter un chiffre d'affaires ou réduire les coûts ou améliorer le fonctionnement d'une chaine de production. Un autre but de cette analyse des données dans le cas d'un commerce consiste à comprendre le comportement d'un client vis-a-vis des achats pour établir de meilleures stratégies de marketing [2].

L'exploration de données appelées aussi datamining est définie aussi comme une technique qui consiste à creuser des informations pour découvrir des exceptions, qui pourraient être classées comme des connaissances précieuses, ce processus et présenté par la figure I.1 [2].

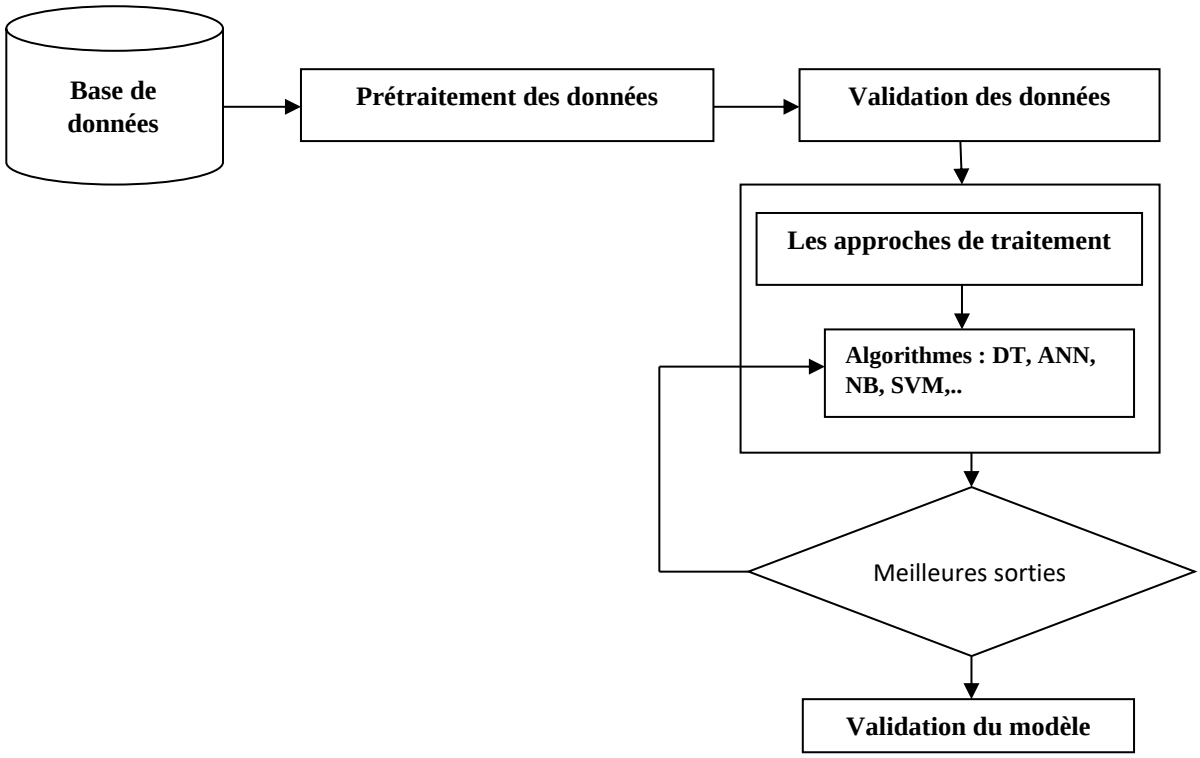


Figure.I.1. Processus du Data Mining

I.2.2 Approches et algorithmes du Data Mining

Le dataminingexplore des algorithmes complexes provenant de plusieurs domaines entre autres l'intelligence artificielle, l'optimisation, résolution des problèmes à satisfcation de contraintes et aussi les statistiques. L'idée fondamentale du datamining repose sur l'utilisation des algorithmes d'apprentissage appliqués sur des données pour l'extraction et la génération de modèles ou patterns [1] [2].

Le datamining possède plusieurs formes de traitement sur les données, nous citons ici les plus exploités dans l'analyse des données et qui sont comme suit [2]:

- R**égression** : le principe de régression est basé sur des variables continues, par exemple : la durée d'hospitalisation d'un patient dans un hopital.

• **Classification** : elles'intéresse à analyser les caractéristiques d'un nouvel objet pour l'affecter à une classe prédéfinie.

• **Prédiction** : les caractéristiques de l'ensemble de données sont classées en fonction du comportement futur ou de la valeur estimée.

• **Clustering** : il s'agit d'une technique de classification non supervisée. Il permet de regrouper les enregistrements en fonction de leurs similitudes de manière que chaque groupe représente un cluster.

Ces aspects de datamining sont réalisés par plusieurs algorithmes dont certains sont plus adaptés à des domaines particuliers tandis que d'autres sont utilisés dans la plupart des problèmes nécessitant une analyse de données.

Nous citons ci-dessous quelques exemples d'algorithmes utilisés dans lalittérature.

I.2.2.1. Régression logistique

La régression logistique (RL) est une technique d'apprentissage automatique elle est utilisée pour les problèmes de classification. La formulede régression linéaire de l'hypothèse est comme suit [3]:

$$h\Theta(x) = \beta_0 + \beta_1 \times X \quad \text{(Eq.1)}$$

$$Z = \beta_0 + \beta_1 \times X \quad \text{(Eq.2)}$$

$$h\Theta(x) = \text{sigmoid}(Z) \quad \text{(Eq.3)}$$

Où :

$$h\Theta(x) = 1/(1 + e^{-(\beta_0 + \beta_1 \times X)}) \quad \text{(Eq.4)}$$

β : les coefficients de régression; Caractéristiques du jeu de données X.

I.2.2.2 Réseau de neurones

Les réseaux de neurones (en Englais Neural Network : NN) sont la technique d'exploration de données la plus utilisée dans les travaux scientifiques. Il s'agit d'une famille d'algorithmes d'apprentissage automatique. Le réseau peut être vu comme un graphe composé de nombreux neurones, chacun d'eux (i) exécute la fonction d'activation fi comme décrit dans l'équation 5 suivante [4]:

$$\tilde{y} = f_i\left(\sum_{i=1}^{n} w_{i,j} \times x_i - b_i\right) \quad \text{(Eq.5)}$$

Où: $\tilde{y}$ est une sortie souhaitée ; w est un poids ; x est une entrée ; b est un biais et n le nombres de caractéristiques.

I.2.2.3 Arbres de décision

Ils sont (en Englais Decision Tree : DT) des algorithmes d'apprentissage supervisé, ils sont très appliqués dans le contexte de la découverte dirigée des connaissances. Ce sont des outils très puissants pour la classification, la description ou l'estimation. Mathématiquement, l'entropie de chaque entité d'un jeu de données est représentée par l'équation 6 [3] :

$$E(x) = \sum_{i=1}^{n} - p_i \log(p_i) \quad \text{(Eq.6)}$$

Où : x : état actuel (nœud), p_i : probabilité d'un événement i d'état x ou pourcentage de classe i dans un nœud d'état x, et n: les étiquettes de classe.

I.2.2.4 Classificateur bayésien

Il est basé sur le théorème de Bayes, c'est une méthode de classificateur statistique utilisée pour prédire les probabilités d'appartenance à une classe. La probabilité qu'un tuple donné appartient à une classe particulière est définie sur la base d'une hypothèse H et d'une évidence E, comme exprimé dans l'équation 7 suivante [5]:

$$P(H/E) = (P(E/H) \times P(H)) / P(E) \quad \text{(Eq.7)}$$

Où : P (H / E) : la probabilité de l'hypothèse après avoir obtenu la prevue

I.2.2.5 Agents intelligents

Les agents intelligents sont des entités logicielles autonomes, dont les versions les plus récentes sont entièrement intégrées dans le processus d'exploration de données. Le comportement des agents intelligents dans le temps peut être formalisé dans l'équation suivante (Eq.8) [6]:

$$\Theta(n + 1) = f(\theta(n), I(n), \mathcal{E}(n)) \quad \text{(Eq.8)}$$

Où : Θ représente la valeur de sortie, f la fonction décrivant les informations transférées à chaque groupe du collectif. Γ: variable d'entrée, elle peut prendre deux valeurs possibles, 0 ou 1, et ε est une erreur entre la sortie souhaitée et une sortie réelle.

I.2.2.6 Voisins les plus proches (K-Nearest Neighbours (K-NN))

C'est l'une des méthodes de classification les plus utilisées, elle appartient au domaine de l'apprentissage automatique supervisé. Ses applications importantes sont la détection d'intrusion, la reconnaissance de formes et l'exploration de données. Cette méthode calcule la distance entre k (caractéristique) et les autres voisins. Par défaut, la fonction k-nn utilise la distance euclidienne ou d'autres métriques pour calculer la distance, elle peut être calculée avec différentes équations mathématiques (distance euclidienne (Eq.9) ; Distance Manhattan (Eq.10) ou Distance Hamming (Eq.11) pour résoudre le problème donné [7].

$$D_e(p, q) = \sqrt{\sum_{i=i}^{n} (p_i - q_i)^2} \quad \text{(Eq.9)}$$

$$D_m(p, q) = \sum_{i=i}^{n} \left| (p_i - q_i)^2 \right| \quad \text{(Eq.10)}$$

$$D_h(p, q) = \sum_{i=i}^{n} \left| (p_i - q_i)^2 \right| \quad \text{(Eq.11)}$$

Où : *p* et *q* sont des sujets ; *n* : fonctionnalités. Le choix de la valeur K à utiliser pour effectuer une prédiction avec K-NN varie en fonction de l'ensemble de données.

I.2.2.7 Règles d'association

Elles reposent sur la syntaxe « if-then » et sont utilisées pour découvrir la probabilité de corrélation entre les différents éléments de données au sein de grands ensembles de données. L'exploration de règles d'association a été utilisée dans diverses applications, telles que la découverte de corrélations de vente dans des données transactionnelles et dans des ensembles de données médicales.

Un exemple d'une règle pourrait être comme suit [8] :

SI ((Age > 60) et (Fumeur = VRAI))
ALOR Risque (crise cardiaque) = 80 % ;
Fin SI

I.2.2.8 Métaheuristiques

Une métaheuristique est une spécification formelle d'un processus de génération itératif qui oriente une heuristique à travers une exploitation intelligente de différents paradigmes pour explorer et exploiter l'espace de recherche. La majorité des cas nécessitent des stratégies d'apprentissage dont l'objectif consisteà structurer l'information pour l découverte des solutions presque optimales. L'une des métaheuristiques les plus populaires est l'algorithme génétique (AG). Le processus génétique comprend les étapes suivantes [9] : la première étape est l'initialisation, puis l'évaluation, après cela, la sélection et la reproduction. Récemment, Cheng et Prayogo [10] ont proposé un nouvelle métaheuristique très prometteur appelé algorithme de recherche d'organismes symbiotiques (SOS : symbiotic organismsearch). L'algorithme SOS simule les stratégies de communication symbiotique que les organismes utilisent pour survivre dans un écosystème. Cependant, l'algorithme SOS ne nécessite pas de paramètres de contrôle, ce qui rend l'algorithme robuste et généralisable contrairement à AG [10][11].

I.2.2.9 Machine à vecteur de soutien (Support Vector Machine : SVM)

Il s'agit d'un type de techniques d'apprentissage automatique, elle permet de résoudre des problèmes de discrimination ou de régression, du fait de leur importante précision des prédictions dans certains problèmes, l'équation 12 suivante exprime un exemple d'une règle utilisée pour résoudre un problème donné [12] :

$$y(x) = w \times x + w_0 \quad \text{(Eq.12)}$$

Où : y : sortie, x : vecteur d'entrée et w : poids.

I.2.2.10 Deep Learning ou apprentissage profond

L'apprentissage en profondeur est considéré comme une nouvelle technique d'apprentissage automatique. Cette technique exploite des réseaux de neurones artificiels dans lesquels plusieurs couches de traitement sont utilisées dans un premier temps, pour extraire les caractéristiques des objets puis l'apprentissage lui-même à partir d'une grande masse de données [13].

I.3 Exploitation de différentes méthodes de datamining en recherche médicale

Nous avons cité ci-dessus dans la précédente section les méthodes et les algorithmes les plus importants et les plus utilisés dans le domaine du datamining, nous dressons maintenant une synthèse bibliographique de quelques travaux cités dans la littérature qui ont exploité ces méthodes dans le domaine de la santé dans les tableaux 1, 2, 3, 4, et 5. Nous avons cité les algorithmes les plus utilisés dans le tableau.6.

Table. I.1. Travaux sur l'application de technique de régression en santé

Références	Objectif	Base de données	L'exactitude	Méthode Data Mining (D.M.)
SALEM et al. [14]	Un système médical de réseaux de zone corporelle sans fil est étudié pour détecter une anomalie pour la surveillance omniprésente des patients et des soins de santé	L'ensemble de données a été collecté l'ECG, la pression artérielle, la respiration, le pouls et l'oxygénation de plusieurs patients cardiaques	96.75%	Machine à vecteur de soutien (SVM)
SENTHILKUMAR et al. [15]	Cette étude vise à fournir une comparaison de la régression logistique avec les techniques d'exploration de données, afin d'identifier les variables prédictives les plus influencées, et aussi de développer un système d'aide à la décision qui permet aux médecins de prendre la bonne décision pour la naissance d'un enfant de faible poids.	Base de données sur les enfants souffrant d'insuffisance pondérale en Inde.	89.95%	Arbre de classification (DT)
DUNCAN et al. [16]	Un modèle de régression a été proposé pour prédire les coûts des soins de santé	Base de données de 30 000 assurés pendant deux ans (2008-2009) des États-Unis	95%	Modèle de regression linéaire
ZAMUDA et al. [17]	Ce travail a mesuré l'importance des fonctionnalités dans l'apprentissage automatique (ML) à partir d'une base de données à grande échelle.	Base de données des maladies rénales chroniques de Slovénie	89.96%	Algorithmegénétique (GA)
Gunasekaran et LOPEZ. [18]	Dans cet article, une architecture de traitement des données de capteurs évolutive a été développée dans le cloud computing, pour mémoriser et traiter les données de capteurs corporels pour les applications de santé.	Base de données sur les maladies cardiaques de Clevelandl'Université de Californie, Irvine (UCI)	RL-SGD81.75%, ANN 89.01%, GA-AIS 87% DT 78.9%	Régression logistique (RL) combiné avec algorithme SGD (descente de gradient stochastique), réseaux de neurones artificiel (ANN), GA-AIS (système immunitaire artificiel), arbre de décision (DT)
GOVENDER et al[19]	Le problème du manque de sang humain en cas de besoin dans les hôpitaux en raison de la pléthore de facteurs externes. Par conséquent, les auteurs ont proposé un modèle mathématique basé sur l'algorithme	6 ensembles de données de banque de sang différents ont été prélevés dans un	SOS 100/98.49% GA 100/76.89%	Recherche d'organismes symbiotiques (SOS), GA, Optimisation de

	SOS pour résoudre ce défi. La stratégie du modèle est la suivante : La banque de sang dispose d'une réserve infinie de capital et d'espace de stockage. il peut importer des modules sanguins avec un approvisionnement illimité ; et tous les groupes sanguins répondent d'abord aux demandes associées à leur groupe sanguin, et à partir de là, le reste de chaque groupe sanguin peut contribuer à d'autres groupes sanguins compatibles	hôpital d'Afrique du Sud	PSO 100/53.94%	l'essaim de particules (PSO)
AYYOUBZADEH[20]	Cette étude a prédit l'incidence de l'épidémie de COVID-19 en Iran.	Jeu de données Covid_19 du 15 février 2020 au 18 mars 2020 en Iran	75.62%	Regression linéaire

Table. I.2. Travaux sur l'application de technique de classification en secteur médical

Références	Objectif	Base de données	L'exactitude	Methode(D.M)
THENMOZHI et al. [21]	Conception et développement d'un système de diagnostic et de prédiction pour les maladies cardiaques basé sur l'exploration prédictive, en utilisant les performances de différentes approches d'exploration de données prédictives telles que l'arbre de décision et les algorithmes Naïve Bayes pour la classification.	Base de données sur les maladies cardiaques (UCI)	DT 96,66 %, NB 99,96 %	DT.; Naïve Bayes (NB)
IYER et al. [22]	Les auteurs ont proposé une technique plus rapide et plus efficace pour diagnostiquer la maladie diabétique et fournir un traitement rapide des patients. Leur méthode analyse les modèles trouvés dans les données grâce à une analyse de classification et utilise des algorithmes d'arbre de décision et Naïve Bayes.	Base de données sur le diabète pour les femmes enceintes (UCI)	DT 76.95 %, NB79.56 %	DT, NB.
AFFONSO et al. . [23]	Présentation d'une nouvelle méthodologie pour classer les images biologiques à l'aide d'un réseau neuronal artificiel rugueux-flou (RFANN). Cette approche a également étudié la sélection des fonctionnalités, afin d'améliorer les performances du processus d'apprentissage par la théorie des ensembles approximatifs (RS).	Ensembles de données sur quelques les maladies (UCI)	L'algorithme ANN est meilleur que la méthode Rugueux-Flou	Rugueux-Flou (Rough-Fuzzy), Réseau de neuronal artificiel (Artificial Neural Networks: ANN)
MANDAL .[24]	Trouve le plus petit sous-ensemble de caractéristiques pouvant garantir une précision de classification élevée du cancer du sein comme bénin ou malin	Ensemble de données sur le cancer du sein (UCI)	NB 94.40%, LR 97.90%, DT 96.50%	NB, LR, DT
NOUREDDINE et al.[25]	La technique métaheuristique SOS a été exploitée dans un processus de classification, la méthode proposée a été appliquée à quatre ensembles de données médicales, dans le but d'aider les experts dans la prise de décision.	Ensembles de données sur quelques maladies (UCI)	SOS 96.5%, GA 81.75%	SOS, GA
OLIVA et al. [26]	Ils ont proposé une nouvelle méthode pour traiter les segments d'électroencéphalogramme (EEG), afin de produire des rapports textuels pour la détection de l'épilepsie.	Base de données de 500 segments ECG.	SVM 95.33%, ANN-MLP 95.09%, 1-NN 86.00%.	DT, 1-voisin le plus proche (1-NN), ANN-MLP, et SVM
ALBAHRI et al. .[27]	Les auteurs ont présenté une approche de sauvetage pour la transfusion du meilleur plasma de convalescence (PC) aux patients les plus critiques atteints de COVID-19.La méthode proposée dépend des exigences biologiques et utilise l'apprentissage automatique et de nouvelles techniques de prise de décision multicritères (MCDM).	Ensemble de données sur le coronavirus Covid-2019	SVM-ML 69%	Méthodes ML (multicouche) pour la classification SVM

Tableau. I.3. Travaux sur l'application de la technique de clustering dans le secteur médical

Références	Objectif	Base de données	L'exactitude	MethodeD.M.
MAHOTO et al. [28]	Dans cet article, les auteurs ont proposé une approche de classification non supervisée, pour extraire des données de santé importantes et nécessaires.	Base de données du diabète	DBSCAN 86.25%, k-means81.40%,	DBSCAN (Density Based Clustering), agglomérathiérarchique et K-moyenne(k-means)
ABDEL-MAKSOUD et al. [29]	Proposition d'une méthodologie pour une meilleure prise de décision dans le traitement des tumeurs cérébrales.	Ensemble de données Web du cerveau.	K-means 95.06% LF 96.05%	K-means et LogiqueFloue(LF)
SHARMILI et al. [30]	L'exploration de modèles de clustering basée sur la neuro floue pour sélectionner les caractéristiques importantes pour prédire les maladies à travers l'ensemble de données GLABRA-180	Ensemble de données GLA-BRA-180	neuro flou 89% NN 68,2%, flou 63,1% KNN 67,3%.	Algorithmes de clustering : neuro flou, NN, flou, KNN
OGBUABOR et al. [31]	Cet article a analysé les performances des techniques de clustering via un ensemble de données de soins de santé, afin de déterminer des algorithmes appropriés qui produisent les clusters de groupe optimisés.	Un ensemble de données sur la maladie a été collecté (UCI)	K-means 97%, DBSCAN 91%	K-moyenne ; Algorithme de clustering DBSCAN
SOHAIL et al. [32]	Les chercheurs ont présenté un travail d'évaluation sur les distributions d'instances et les groupes de grappes pour la classification des diagnostics épidémiologiques des maladies. Ils ont utilisé des données manuelles extraites de nombreux référentiels, afin d'évaluer les points euclidiens.	Ensemble de données sur les diabétiques	K-means 50% SOM60%	La K-moyenne et la carte auto-organisée (SOM)
NOUREDDINE et al.[33]	Ils ont utilisé la méthode de regroupement en utilisant l'algorithme métaheuristique de recherche d'organismes symbiotiques (SOS) pour détecter s'il y a une maladie ou non.	Ensembles de données (UCI) sur 1-le cœur, 2-le foie, 3-le diabète et 4-le cancer du sein	SOS :63.75%, 61.57%, 60,10%, 96.48%/ GA : 65.25%, 52.75%,39.32%? 96.50%/	Algorithmemétaheuristique SOS, GA
SHALABY et al. [34]	La maladie de Parkinson (MP) est une maladie neurodégénérative clinique présentant des symptômes tels que tremblements, rigidité et incapacité posturale. Selon Harvard, environ 60 000 citoyens américains reçoivent un diagnostic de MP chaque année. Le diagnostic de la MP repose sur l'expertise du médecin et dépend de plusieurs critères cliniques établis. Cela rend le diagnostic subjectif et inefficace. Par conséquent, des efforts continus sont déployés pour améliorer le diagnostic de la MP en utilisant des approches d'apprentissage en profondeur qui s'appuient sur des neurologues expérimentés. Pour cette raison, les auteurs ont utilisé un réseau de neurones siamois basé sur le clustering pour diagnostiquer la MP	Ensembles de données sur La maladie de Parkinson	84.02%	Réseaux de neurones siamois (SNN)

Tableau. I.4. Travaux sur l'application de la technique de prédiction dans le domaine de la santé

References	Objective	Base de données	L'exactitude	D.M. Method
JAJROUDI et al. [35]	Les auteurs ont étudié l'importance de la prédiction de survie pour les médecins et les patients, afin de trouver la meilleure façon de traiter les patients atteints de cancer de la thyroïde. Ils ont également proposé une méthode prédictive pour les aider.	Ensemble de données sur le cancer de la thyroïde	LR 86.8% ANN-MLP 96%	ANN-MLP, et LR
JOSHI, et al. [36]	L'utilisation des connaissances découvertes peut aider les cliniciens à diagnostiquer les maladies cardiaques et à prédire les conditions à risqué.	Base de données sur les maladies cardiaques (UCI)	DT 92,2% BN 84,2% KNN 100%	D.T., K-voisins les plus proches (KNN), et NB
THAPEN, et al. [37]	Présentation de l'approche DEFENDER, un système logiciel qui intègre les données collectées à partir des médias sociaux et des actualités, et utilise des algorithmes pour détecter les flambées épidémiques, la connaissance de la situation, la prédiction et le prétraitement des risques.	Base de données de Twiter des patients	37%	Leréseau de localisation. DEFENDER
RAHIM, et al. [38]	Le virus de la dengue est mortel et il est transporté par les moustiques jusqu'aux habitants de Malaisie. Pour enrichir cet objectif, les chercheurs ont présenté une approche pour prédire les cas de dengue à l'aide de l'exploration de données.	Ensemble de données sur la dengue en Malaisie	NB 72%, DT 94%, LR 93%, SVM 93%, RF 95%	NB, DT, LR, SVM, et Forêtaléatoire (RF).
NILASHI, et al. [39]	Un algorithme puissant pour le diagnostic de la maladie de l'hépatite est proposé, en exploitant les avantages de l'apprentissage d'ensemble.	Ensemble de données sur l'hépatite de l'UCI	K-NN 71.41%, SVM 81.17%, ANN 78.31%, ANFIS 79.67%, NIPALS–SOM–ANFIS ensemble 93.06%, PCA–LSSVM95.00%, PCA–AIRS94.12%, LFDA–SVM96.77%, RES–ELM100%	K-NN, SVM, ANN, ANFIS, NIPALS–SOM–ANFIS ensemble, PCA–LSSVM,PCA–AIRS, LFDA–SVM, RES–ELM
SAJEEV et al. [40]	Cette étude a discuté des variations qui peuvent survenir dans les performances de certaines méthodes de prédiction d'apprentissage automatique non linéaires et plus sophistiquées typiques sur une étude de cas de maladie cardiaque.	Ensemble de données sur les maladies cardiaques (UCI)	DL94% LR84.8% SVM88.8% RF93.3%	Deep learning (DL), LR , SVM, et Forêtaléatoire(Random forest: RF)
ALBAHRI et al. [41]	Une application basée sur une technique d'exploration de données et des algorithmes d'apprentissage automatique (ML) est présentée, pour détecter et diagnostiquer l'épidémie de COVID-19.	Ensemble de données Covid_19	SVM 86.44% ANN 90% NB 73.6% DT 71.58%.	DT (j48), SVM, ANN, et NB.
AYON et al. [42]	Les maladies cardiaques s'avèrent être la aladie la plus courante et l'impact de cette aladie est plus dangereux que toutes les autres aladies. Ainsi, les auteurs ont comparé un rtain nombre de techniques d'intelligence ificielle pour la prédiction de la maladie ronarienne.	Deux ensembles de données sur les maladies cardiaques : Cleveland et Hongrois de (UCI)	SVM 69.8%/84.4%, LR 70.3%/92.2%, ANN-MLP 81.1%/89.3%, RF 70.7%/87.3% , DT 75.4%/77.6%, NB82.4%/83.4%	SVM, LR, ANN-MLP, RF, DT, et NB.
ZabirulIslam et al. [43]	Ils ont proposé un modèle atteint une précision de détection élevée de COVID-19	COVID-19	94%	Réseau de neurones convolutifs (CNN)
MOHMMADZADEH et al.[44]	L'article a présenté trois approches binaires basées sur SOS pour résoudre le problème de sélection de caractéristiques.	Ensembles de données de (UCI) : 1-cancer du sein, 2-patiente indienne du foie, 3-hépatite 4-cancer du poumon	SOS97.46%- 75%- 79.5%- 100% GA 97.1%- 72.9%- 69.2%- 93.8% PSO 97.1 %- 71.8 %- 67.7 %- 90.6 %	SOS, GA, PSO

Tableau. I.5. Travaux sur l'application de deux techniques ensemble de la prédiction et de la classification en santé

Références	Objectif	Base de données	L'exactitude	MethodeD.M.
METHAIL A I et al. [45]	Techniques de modélisation de classification d'exploration de données utilisées dans la prévision des maladies cardiaques.	Base de données sur les maladies cardiaques de Cleveland (UCI)	NB 86.53 % DT 89% NN 85.53%	Arbres de décision, Naïve Bayes et réseau de neurones.
VIJAYARA NI et al. [46]	Ils ont utilisé des algorithmes d'exploration de données de classification pour prédire la maladie rénale.	Ensemble de données sur les maladies rénales (UCI)	NB 70.96% SVM 76.32%	SVM et Naïve Bayes
WU, et al. [47]	Les auteurs ont proposé un algorithme SOS combiné à ANN pour prouver l'efficacité du SOS pour résoudre les problèmes de classification et de prédiction dans différents ensembles de données.	Ensembles de données (UCI)de : 1-sang, 2-troubles hépatiques 3-statlog (Cœur)	SOS 82.74%- 75.42%- 85.86% GA 76.86% - 55.08%- 70.65% PSO 80.39%- 72.03%- 86.95%	SOS-ANN, GA-ANN, et PSO-ANN
BAITHAR U et al. [48]	Ils ont utilisé différents algorithmes de classification pour prévoir les maladies du foie, notamment les voies biliaires, le cancer du foie, la cirrhose, l'hépatite chronique et l'hépatite aiguë à partir du test de la fonction hépatique (LFT).	Ensemble de données sur les troubles hépatiques (UCI)	J48 68.97% ZeroR 57.971% MLP 71.59% 1BK 62.89% NB 55.36% VFI 60.28%	D.T (J48)., NB, MLP, ZeroR, K-NN (connu sous le nom IBK) et algorithme VFI.
JALALI, et al.[49]	Cet article présente une approche d'exploration de données pour détecter le cancer du sein, en utilisant des règles d'association et à travers des données réelles, obtenues auprès de l'un des hôpitaux iraniens.	L'ensemble de données sur le cancer du sein a été compilé à partir de l'hôpital Neyriz en Iran.	DT 78.03% . SVM 93.72% K-NN 87.73% NB 77.21% RF 91.21% NN 92.33%	DT , SVM, K-NN, Naïve Bayes, Forêt aléatoire(RF), et Réseaux de neurones (NN).
ALJAWAR NEH,et al.[50]	Une fonction de distance nommée MANTRA a été proposée pour imputer les valeurs de données manquantes dans le domaine médical.	Ensembles de données biomédicales	99.71%	Approche MANTRA
AYON et al. [51]	Ils ont proposé une stratégie pour le diagnostic du diabète à l'aide d'un réseau de neurones profonds	L'ensemble de données sur le diabète indien Pima (PID).(UCI)	98.35%	Deep Learning
SORNALA KSHMI, et al. [52]	Les auteurs ont analysé les différentes perspectives et connexions entre des informations apparemment sans rapport, pour prédire les tendances possibles de l'état de santé et du comportement du patient.	Base de données Physionet	ARM 96% EAA-SMO 98%	Association de règles d'exploration (ARM), Algorithme Apriori amélioré (EAA), SMO et méthode hybride EAA-SMO

À partir des tableaux I :1, 2, 3, 4 et 5, ces techniques ont été exploitées pour obtenir un avis précoce de la découverte d'une épidémie, la prédiction des niveaux de maladie et la capacité d'estimer la possibilité de symptômes chez les personnes. Bien que ces techniques soient des outils pratiques, mais certaines d'entre elles sont puissantes comme NB, DT, LR, ANN-MLP, SVM, Deep Learning et SOS qui donnent de meilleurs résultats (entre 80% et plus de 90% en termes del'exactitude). Une partie de notre contribution est resumé dans le tableau I.6 qui synthétise les algorithmes les plus couramment utilisés en termes de précision, de facilité d'utilisation, de complexité et d'applicabilité.

Tableau. I.6. Description des algorithmes les plus utilisés

Algorithme	Intervalle de Précision	Complexité & Remarque
SVM	76.32% - 95.33% Bon	• Se comporte bien en haute dimension avec une lente mise œuvre. • C'est un bon algorithme dans le cas où les caractéristiques sont séparables. • SVM peut éviter les anomalies. • Mauvaises performances pour les données complexes
ANN	90% - 96% Bon	• Capable d'apprendre et de modéliser des bases de données complexes et bruyantes. • Peut compenser les données incomplètes. • Possibilité de traitement parallèle ce qui entraine une consommation élevée du temps et du matériel.
DT	76.95 %-96,66 % Bon	• Excellente performance pour une masse de données. • Peut gérer des données manquantes. • Sélectionne automatiquement les fonctionnalités • Les arbres plus grands produisent un encombrement de mémoire.
NB	70.96% -97,96 % Bon	• Utilisable en temps réel • Fonctionne bien avec les masses des données, les multiclasses, et les données non pertinentes • Mauvaise prédiction • Les données d'entraînement doivent bien représenter la population.
K-means	81.40% - 97% Bon	• Il est facile à implémenter • Capable de gérer des données volumineuses. • Sensible aux données bruitées ou manquantes
LR	86.8%-97.90% Bon	• Facile à mettre en œuvre et efficace. • Peut traiter des fonctionnalités quelle que soit leurs tailles. • Incapable de manipuler les données non linéaires.
K-NN	71.41%-87.73% Bon	• Simple de mettre en œuvre ; • Peut résoudre des problèmes multi-classes. • Incapable de gérer les fonctionnalités non pertinentes et les problèmes complexes. • Prend un long temps d'exécution
RF	91.21%-95% Bon	• Utile pour les données en masse • Performant sur les données non pertinentes et incomplètes. • Ses paramètres sont inaccessibles. • C'est long à mettre en œuvre.
GA	55.08%- 97.1% Bon	• Efficace et robuste pour diverses classes de problèmes d'optimisation. • Peut-être piégé dans des optima locaux. • Complexité de traitement élevée, • Lent temps de convergence

SOS	75%-100% Très bon	• Robuste et stable pour les problèmes d'optimisation et avec des solutions de haute qualité. • C'est un algorithme simple et facile à comprendre et à mettre en œuvre, • Temps d'exécution raisonnable. • Peut gérer une masse de données et converge rapidement vers l'optimum global. Faibles performances pour les problèmes d'optimisation complexes
Deep Learning	84.02% - 98.35% Excellent	• Excellente performance pour les Bigdata quels que soient leurs types. • Extraction automatique des fonctionnalités • Nécessite un coût élevé.

Nous signalons dans cette section qu'à partir de notre étude de synthese que pour l'analyse de données et pour une optimalitée du résultat, deux critères s'imposent inevitablementqui sont : le choix de la technique appropriée et le type de données à traiter.

Il est donc important de choisir la technique adéquate qu'on peut utiliser. De plus, pour sélectionner la technique d'exploration de données appropriée, nous devons prendre en considération la simplicité de la méthode et son efficacité pour résoudre des problèmes combinatoires complexes. Ces techniques doivent être capables de gérer les problèmes des bases de données tels que la dynamique ; l'incompletude ; les bruit et le nombre élevé des caractéristiques.

En conséquence, le processus d'exploration de données est un outil puissant permettant aux experts de prendre leurs décisions avec précision. Les améliorations continuent de ces processus augmentent considérablement la fonctionnalité des différents types des problèmes.

I.4. Domaine d'application du Data Mining

De nos jours, le datamining est devenu un refuge inévitable pour les secteurs socio-économiques manipulant une masse de données volumineuse. L'apport du datamining en matière de gain de temps, d'extraction de connaissances très onéreuses en matière de prise de décision a poussé son utilisation dans plusieurs domaines.

Dans cette section, nous dréssons une liste de quelques imporantes applications d'exploration de données.

I.4.1. Finance

Dans le secteur bancaire et financier, la gestion qualitative s'appuie sur des données fiables et de haute qualité. Une tache bancaire de grande nécessité consiste en un suivi rigoureux des marchés des prêts, les données financières et des utilisateurs peuvent à diverses fins telles que la prévision des remboursements de prêts et la détermination des intérets de crédit. C'est ainsi que les techniques d'exploration de données rendent ces tâches plus faciles à gérer [1].

I.4.2. Télécommunications

Le domaine de télécommunication est devenu un environnement de stockage et de traitement des données ayant un grand volume et de grande importance suite à l'abonnement des millions d'utilisateurs au réseau téléphonique.

Les fournisseurs de services désirent connaitre parfois en temps réel certaines propriétes des abonnés et l'utilisation des techniques d'analyse de données et d'extraction de connaissances sous forme de modèles et des statistiques pour une décision de management est le seul moyen pour y parvenir [1].

I.4.3. Enseignement supérieur

Dans ce contexte de manipulation de données, les établissements universitaires recherchent des solutions innovantes pour répondre aux besoins croissants dans certaine filières. Ils peuvent exploiter les données des étudiants pour prédire quel étudiant s'inscrirait à une formation en se basant par exemple sur la moyenne obtenue au baccalauréat ainsi que sa spécialite cela permettra d'avoir une vision sur le nombre d'étudiant par filière etc [1].

I.4.4. Surveillance d'entreprise

La surveillance d'entreprise est le domaine qui s'interesse à la surveillance du comportement d'une personne ou d'un groupe de personne par une entreprise. Il sagit de collecter des données qui par lasuite sont généralement utilisées à des fins de marketing ou parfois ces mêmes données sont vendues à d'autres sociétés, mais sont également régulièrement partagées avec des organismes gouvernementaux. Les données sont par exemple utilisées dans le domaine du marketing comme les publicités sur Google, facebook et Yahoo etc [1].

I.4.5. Industrie

Dans les environnements de fabrication modernes, de grandes quantités de données sont collectées dans les systèmes de gestion de bases de données et les entrepôts de données de tous les domaines concernés, y compris la conception de produits et de processus, l'assemblage, la planification des matériaux, le contrôle de la qualité, la maintenance, la détection des défauts, etc [1].

Les entreprises de fabrication utilisent depuis plusieurs années différentes types de techniques et d'outils pour atteindre leurs objectifs fondamentaux. Parmis ces techniques, nous citons les techniques et outils d'apprentissage automatique (Machine Learning) et d'exploration de données (DataMining) qui sont très utile pour relever les défis de la fabrication [1].

I.4.6. Aviation

Le domaine de la sécurité opérationnelle et de la qualité de l'aviation a connu un intérêt et une utilisation accrus des techniques d'exploration de données pour l'analyse de la sécurité, l'examen des incidents, la détection des pannes et les améliorations de l'efficacité [1].

L'avion est devenu un véritable objet connecté à cause de l'utilisation de capteurs sur les nombreux équipements embarqués. Les experts en aéronotiqueseront capables d'établir un suivi rigoureux de l'état des équipements. En conséquence, toute information sur une hypothétique défaillance technique est automatiquement détectée par le système. Ils pourront ainsi traiter l'information reçue et prévoir les actions de maintenance nécessaires, c'est ce qu'on appelle la maintenance prédictive [1].

L'analyse des données de vol obtenues à partir des enregistreurs de bord permet d'identifier les tendances et les modèles d'exploitation et les occurrences d'événements critiques pour la sécurité. La surveillance des données de vol est régulièrement mise en œuvre aussi dans les opérations commerciales où l'application de techniques d'exploration de données a été proposée et démontrée pour la détection d'anomalies [1].

I.4.7. Transport

Le transport est devenu en espace générateur de grand volume de données suite à l'évolution du matériel exploitant les technologies les plus avancées. Nous citons ici deux applicatons qui ont réalisés des études d'analyses de données pour le transport.

Dans le premier projet, les auteurs ont exploré l'effet des capacités de la chaîne d'approvisionnement basées sur les données sur le transport (basé sur le train) pour le calcul du chemin le plus court. Pour cela, ils ont exploité les données open source de « London Underground Transportation » par une analyse en mettant en œuvre différents outils d'exploration de données [53].

Dans le second projet, les auteurs se sont interessés à l'exploitation du datamining pour le transport urbain suite au constat de l'importance de cette activité et de la concurrence entre les entreprises activant dans ce domaine. L'interet s'est porté sur l'utilisation de plusieurs techniques comme les règles d'association et de classification, l'algorithme K-means ainsi que la classification Naive Bayes (probabiliste) pour étudier les facteurs affectant le coût et le temps du transport routier et ferrovier [53].

I.4.8. Agriculture

Les problèmes dans le domaine agricole peuvent être efficacement résolus en utilisant des techniques d'exploration de données, car elles sont anticipées à l'aide de données brutes.

L'agriculture intelligente est l'une des solutions pour faire face à la demande croissante de nourriture tout en répondant aux exigences de durabilité.

Dans ce contexte de recherche, les données ayant un volume assez important sont la pierre angulaire de réussite de toute activité de calcul ou de prise de décision. Ces données concernent les conditions météorologiques, les sols, les maladies, les insectes, les semences, les engrais, etc. Leprocessus intelligent consiste à collecter, transmettre, trier et analyser des données [1].

I.4.9. Analyse des données biologiques et médicales

L'exploration de données offre un grand potentiel pour améliorer les systèmes de santé. Elle manipule des données et à travers des analyses pour identifier les meilleures pratiques qui améliorent les soins et réduisent les coûts. Les chercheurs utilisent des approches d'exploration de données telles que les bases de données multidimensionnelles, l'apprentissage automatique, la visualisation de données et les statistiques. L'exploitation minière peut être utilisée pour prédire le volume de patients dans chaque catégorie. Des processus sont développés pour s'assurer que les patients reçoivent les soins appropriés au bon endroit et au bon moment, ce processus est appelé « Informatique Ubiquitaire » qui tient compte de la position géographique, du profil du patient, des outils d'accès à distance ainsi que le temps. L'exploration de données peut également aider les assureurs-maladie à détecter les fraudes et les abus [1].

Nous notons que l'avancé des technologies de l'information et decommunication a largement influencée la télémédecine. L'utilisation de l'internet des objects, les capteurs mobiles, les réseaux de communication 4 et 5 G, la télémedecine est devenue une forme économique et d'aide à la décision prédictive médicale en temps réel très demandé dans la pratique et dans le monde. Certe le côut d'une solution de télémdecine et parfois très grand mais l'apport en matière d'assitancemédiclae à distance lui donne une place primordiale pour certaine application dans la médecine.

I.4.10. Autres applications scientifiques

Le datamining possède d'autres aspects applicatifs à travers des simulations numériques rapides dans des domaines scientifiques manipulant un volume de données assez important comme le génie chimique, la dynamique des fluides, le climat et la modélisation des écosystèmes qui génèrent de vastes ensembles de données. L'exploration de données produit des fonctionnalités telles que les entrepôts de données, le prétraitement des données, la visualisation, l'exploration basée sur des graphes, etc [1].

I.5. Télémedecine et travaux connexes

Dans cette section, nous donnons la définition de la télémédecine ainsi que les actes adoptés pour son application.

I.5.1 Définition de la télémedecine

La télémédecine, également appelée télésanté ou e-médecine, est lexercice à distance de services de santé, y compris des examens et des consultationsà travers un processus de télécommunication. La télémédecine permet aux acteurs médicaux d'évaluer, de diagnostiquer et de traiter les patients sans avoir besoin d'une visite en personne. Les patients peuvent interagir avec les médecins depuis leur domicile en exploitanat les outils technologiques personnelle [54]. Pour les patients à domicile, un examen de télémédecine est basé sur le partage des informations sur les antécédents médicaux et les symptômes par le patient avec l'expert médical. Sur la base de l'évaluation du clinicien, et suite à cette interaction le patient recevra des instructions supplémentaires, telles que prendre des médicaments en vente libre, se rendre à l'hôpital ou prendre un rendez-vous de suivi [54].

I.5.2 Actes de télémédecine

La télémédecine possède plusieurs aspects de ses pratiques médicales, ils sont précisés dans la figure I.2 [55] :

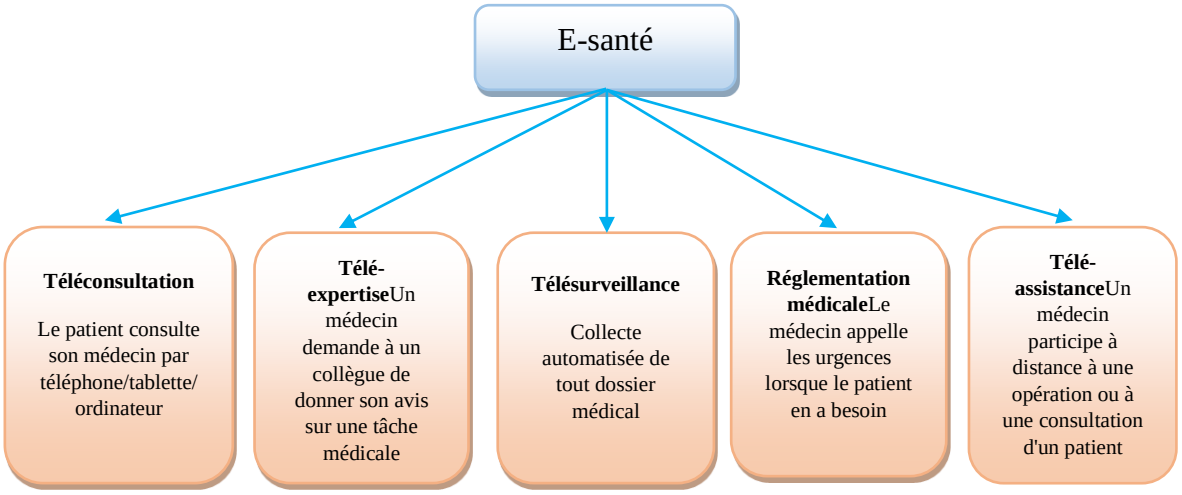


Figure I.2. Les cinq actes de télémédecine

• ***Téléconsultation*** : la téléconsultation est une consultation réalisée à distance par un médecin généraliste ou spécialiste par transmission vidéo avec le patient. Lors de cette téléconsultation, le patient peut être à la maison, seul ou accompagné d'un autre professionnel de santé. Il discute avec le médecin via un ordinateur, une tablette ou un smartphone équipé d'une webcam et connecté à internet. Le patient se place dans un environnement à proximité de son domicile [55].

• ***Télé-expertise*** : c'est la tâche qui rassemble deux ou plusieurs professionnels de la santé pour exprimer à distance leur avis d'expert sur le dossier médical d'un patient. La télé-expertise permet à un médecin de demander l'avis d'un collègue sur la prise en charge d'un patient via une messagerie sécurisée ou une solution de télémédecine [55].

• ***Télésurveillance médicale*** : cette approche permet de poursuivre un patient souffrant d'une maladie chronique. Le patient est suivi à domicile par des indicateurs cliniques et / ou biologiques choisis par un médecin de santé. Les données sont collectées volontairement par un dispositif médical à l'aide d'algorithmes construits pour la pathologie concernée ou saisies directement par le patient, puis transférées au professionnel de la santé via des services de télésurveillance commerciale [55].

• ***Télé-assistance médicale*** : cette action donne à un professionnel de la santé la possibilité d'assister à distance un professionnel de santé non médical [55].

• ***La réponse médicale apportée dans le cadre de la réglementation médicale*** : elle se présente sous la forme de plusieurs services médicaux, dont l'aide médicale d'urgence et le conseil médical personnalisé sont les formes les plus utilisées [55].

I.5.3. Impact du datamining en télémédecine

Les techniques de datamining peuvent aider les médecins à prendre des décisions médicales en identifiant les diagnostiques. De plus, en analysant le profil du patient, ses antécédents, son diagnostic et les traitements antérieurs, de nouveaux plans de traitement peuvent être suggérés à l'aide du processus d'exploration de données (datamining), qui est illustré dans la figure I.3 [56]. D'autre part, la télémédecine présente de nouveaux avantages pour l'intégration de l'exploration de données dans les systèmes d'information sanitaire, car elle augmente le volume de données collectées dans tous les domaines clés pour lesquels l'exploration de données est appliquée avec succès [56].

Les programmes de télémédecine produisent beaucoup plus de données diagnostiques que les diagnostiques eux-mêmes. L'analyse et l'exploration des données effectuées sur ces données peuvent non seulement révéler de nouvelles informations sur les maladies diagnostiquées, mais également optimiser d'avantage les paramètres des tests de diagnostique appliqués [56]. Par conséquent, l'exploration de données est d'une grande importance pour le domaine de la santé, car les connaissances acquises aident à mieux comprendre les besoins des organisations de soins de santé et à prendre des décisions qui contribueraient à améliorer la santé des patients.

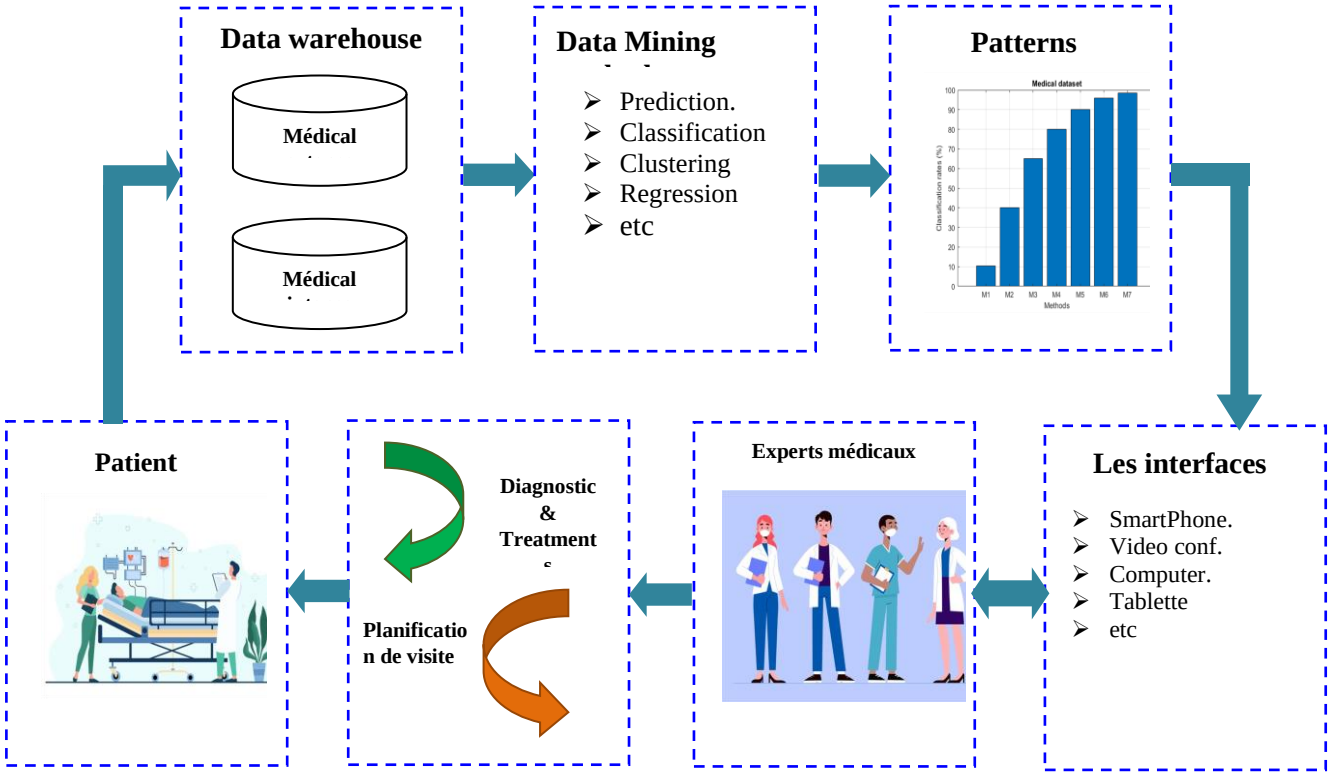


Figure. I.3. Processus du Data Mining dans le secteur médical

D'après la figure I.3, pour chaque application de ce processus, le choix d'une méthode de datamining dépend des objectifs de l'utilisateur, cela dépend aussi du type de données et surtout de la nature du problème que nous devons résoudre. Certains problèmes nécessitent la combinaison de deux ou plusieurs techniques d'exploration de données qui peuvent être séquentielles ou parallèles pour obtenir d'excellents résultats.

Les statistiques [57][58] présentées dans les figures I.4 et I.5 indiquent que le domaine de la santé est en croissance, ce qui pourrait être dû aux facteurs suivants [57][58] :

• ***Flexibilité*** : de cette manière, les ressources de soins de santé sont accessibles depuis n'importe quel environnement (rural ou urbain), à tout moment.

• ***Accessibilité*** : les applications de télémédecine sont accessibles depuis n'importe quel environnement et donnent accès à des nombreuses ressources de santé.

• ***Portabilité*** : il s'agit de l'adaptabilité et de la coexistence de toutes les sources courantes de périphériques informatiques avec l'application installée.

• ***Réduction du temps*** : étant donné que la technologie prend du temps et utilise de l'espace, les médecins et les patients n'ont pas besoin d'être au même endroit en même temps ; ils peuvent simplement communiquer en temps réel via les applications sur le dispositif informatique.

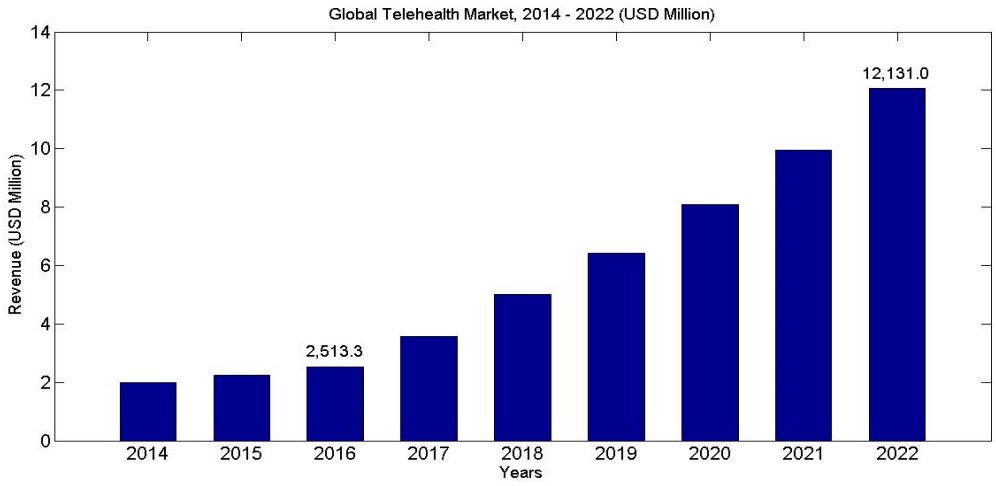


Figure. I.4. Evaluation du marché de la télémédecine pendant la période 2014-2022 [57]

Les applications de télémédecine les plus courantes sont illustrées ci-dessous :

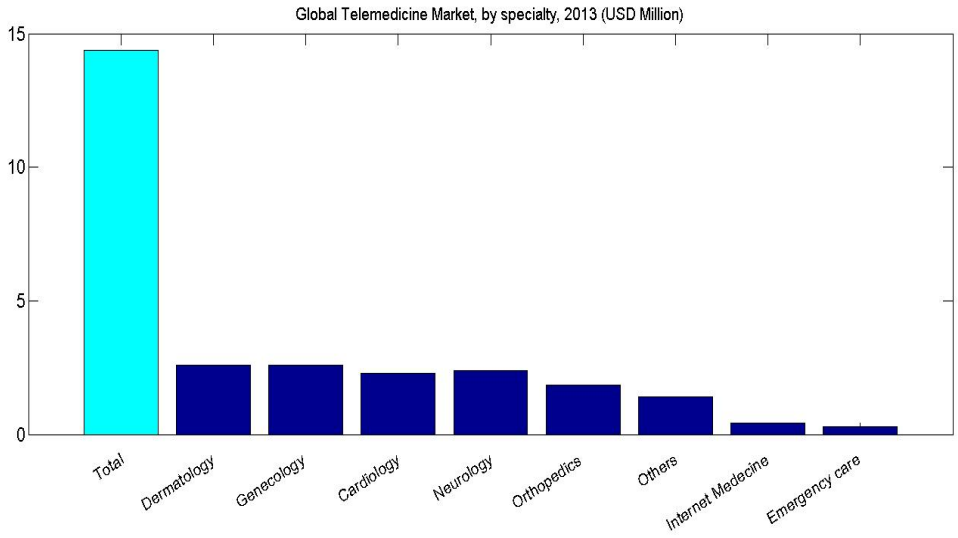


Figure. I.5. Catégories d'applications de télémédecine [58]

D'après les figures I.4 et I.5, force est de constater que parmi les nombreuses recherches et projets de télémédecine qui ont été identifiés, seuls quelques-uns ont été mis en œuvre en 2013 en raison de plusieurs facteurs dont : son coût élevé, et la méconnaissance de son utilisation par de nombreuses personnes, notamment les personnes âgéesetc. Cependant, il existe certaines institutions de santé comme : Partners Health Care aux États-Unis ; TeladocInc, qui supervise la prise en charge de son application et a obtenu un grand succès, notamment pour les maladies chroniques qui nécessitent un suivi continu. La figure I.5 montre le pourcentage du nombre de maladies auxquelles la télémédecine est appliquée, telles que : les maladies cardiaques, la dermatologie, la neurologie, l'orthopédie, les soins d'urgence, la médecine interne, la gynécologie, autres.

Les régions où la télémédecine est largement appliquée sont : les États-Unis, le Brésil, le Mexique, le Canada, l'Allemagne, le Royaume-Uni, la France, l'Italie, l'Espagne, la Chine, l'Inde, le Japon, l'Australie, la Nouvelle-Zélande, le Moyen-Orient et l'Afrique (les pays du Sud Afrique) [58].

I.5.4. Travaux du Data Mining réalisés en télémédecine

Dans la littérature, plusieurs techniques sont utilisées dans le traitement des données médicales par datamining. Nous citons des applications courantes et les plus utilisées dans le tableau I.7 ainsi que quelques travaux de recherche sur l'exploration de données en télémédecine présentés dans le tableau I.8.

I.5.4.1 Applications de télémédecine les plus utilisées

La technologie de télémédecine développée sous forme d'assistance médicale et de consultation par vidéo, message, téléphone, chat, etc [59] a touché plusieurs spécialités avec des approches différentes, qui est un processus d'analyse de données volumineuses sous différents angles afin d'identifier des tendances ou des modèles menant à une intelligence stratégique médicale. Nous citons dans le tableau I.7 les applications les plus connues et les plus utilisées en 2020 [59].

Tableau I.7. Les applications de télémédecine les plus utilisées en 2020

Nom & référence de l'application	Objective
AmWell [60]	L'application fournit une assistance médicale 24h/24 et 7j/7 pour les soins d'urgence, la thérapie, la psychiatrie et d'autres situations à un coût fixe.
MDLive [61]	Il s'agit d'une plate-forme sur laquelle les patients peuvent contacter des médecins, des conseillers, des dermatologues ou des psychiatres pour effectuer des consultations au moyen d'une application mobile à un coût raisonnable.
Dialogue [62]	Cette application est destinée aux employés canadiens dans le domaine des affaires. Il propose aux infirmières des consultations en utilisant la vidéo en ligne et la livraison de médicaments en temps réel, c'est gratuit.
Doctor On Demand [63]	Cette application de haute qualité fournit des consultations médicales en utilisant un système d'apprentissage en profondeur et oriente les patients vers des experts et des services de santé aux États-Unis, mais avec un coût élevé.
Lemonaid [64]	Il offre certains services tels que: la prise de rendez-vous chez le médecin sur le cholestérol, les sinus, le reflux acide et les tests de laboratoire. C'est une application mobile fiable à un coût fixe.
Doxy.me [65]	Doxy.me propose des consultations gratuites à tous les professionnels de santé avec leurs patients.
Endoluxe [66]	Cette application connecte tous les services hospitaliers aux médecins pour suivre l'état du patient. C'estunegrande innovation mais facile à utiliser.
PlushCare [67]	Soins d'urgence virtuels dans le traitement de la santé mentale. La possibilité de parler aux meilleurs médecins de haute qualité à l'aide d'un smartphone à un coût fixe.

Simple Contacts [68]	Une application simple qui aide les patients à renouveler leurs lentilles de contact à l'aide d'un smartphone sans avoir besoin d'aller chez l'opticien pour un prix raisonnable pour télécharger cette application.
Kinesiometrics [69]	Il relie les centres médicaux universitaires, les dispositifs médicaux, les médecins et les compagnies d'assurance pour suivre les patients après la chirurgie afin d'étudier l'effet de l'activité physique sur les résultats des patients afin d'aider à la prise de décision concernant l'état du patient.
First Opinion [70]	Cette application aide le patient à connaître le premier avis à travers les symptômes de la maladie en communiquant avec les médecins pour obtenir leurs conseils à l'aide de messages gratuits.
ThriveHealth [71]	Cette application permet une visite virtuelle ou une visioconférence pour les patients et leurs familles. En plus de pouvoir consulter des médecins. Il offre des soins de haute qualité.
Dentulu [72]	C'est une application mobile qui permet au patient de prendre rendez-vous avec le dentiste.

Le tableau I.7 représente les meilleures applications utilitaires en ligne. Il est intéressant de connaître la nécessité d'analyser ce grand volume de données médicales et d'en extraire des informations utiles telles que la prédiction de diagnostic, la planification de visite, le partage d'expertise médicale, la distribution de médicaments, etc. Ces applications reposent sur plusieurs approches de datamining telles que : ANN, SVM, DT, Deep Learning, etc. Ils sont bons en termes de qualité, de facilité d'utilisation, d'étendue de la crédibilité et de coût raisonnable. L'utilisation des applications de la télémédecine augmente, surtout après l'émergence de COVID-19. On constate que l'usage de la télémédecine et les restrictions sanitaires notamment le confinementont imposé un mode de vie très complexe qui a clairement affecté l'usage de la télémédecine. Ainsi, les téléconsultations sont passées d'un usage épisodique à un usage quasi universel, c'est-à-dire que toutes les maladies et types de patients ont été pris en compte [73].

I.5.4.2 Travaux de recherche basés sur une approche de Data Mining en télémédecine

Les résultats de l'étude indiquent que la télémédecine améliore le diagnostique simplifie le processus de transport du patient et raccourcit le séjour. La télémédecine a eu un impact sur les décès et les complications chez les patients qui ont des maladés chroniques, contagieux ou traumatisés dans les zones rurales, surtout au cours des six dernières années (2014-2020) avec le développement de dispositifs de communication, comme le montrent les figures I.6 et I.7.

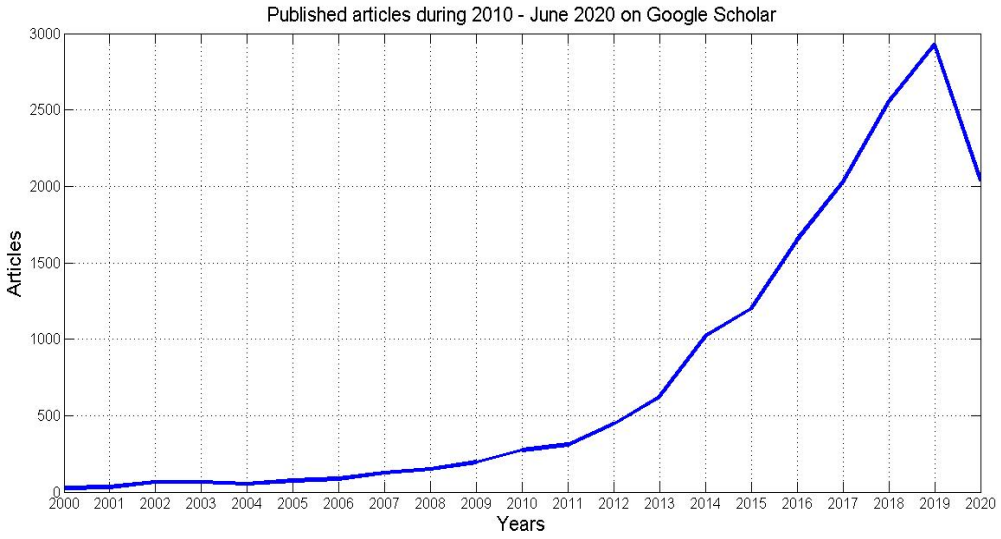


Figure. I.6 Courbes des articles publiés de 2000 à juin 2020 sur Google Scholar

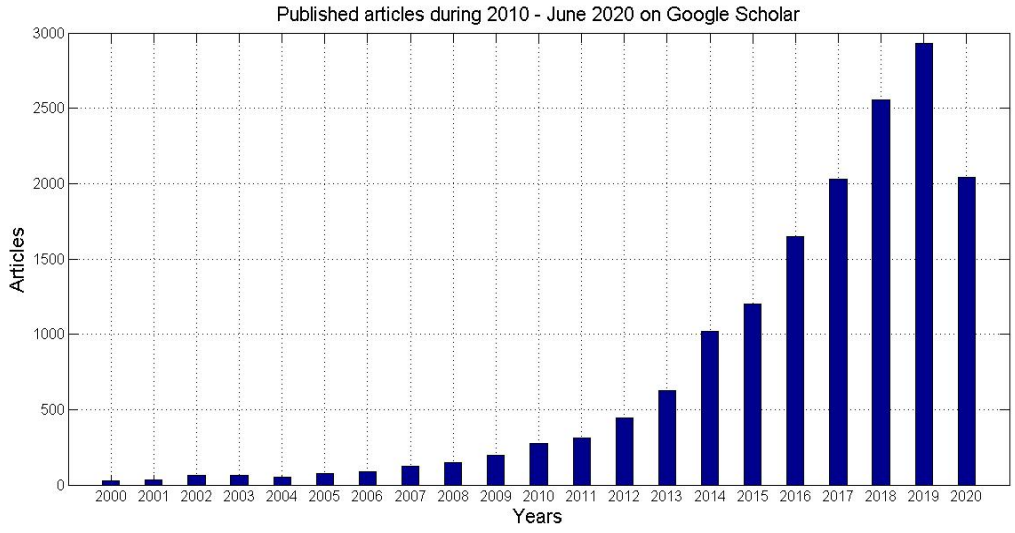


Figure. I.7. Histogramme des articles publiés entre 2000 et juin 2020 sur Google Scholar

A partir des figures I.6 etI.7 une baisse significative en 2000-2010. Google Scholar n'a cité que 400 articles par rapport à l'évolution des dispositifs de communication. De 2010 à 2014, le nombre d'articles publiés a légèrement augmenté pour atteindre 1000 articles. En 2014-2020 ; le nombre d'articles a fortement augmenté d'environ 10000, et a également augmenté de manière significative à cause de son application réelle dans le domaine médécal.

Le nombre plus élevé d'articles publiés conduit à une amélioration importante des applications de télémédecine dans le secteur médical. Il a également augmenté de manière significative son application réelle dans le secteur de la santé. Ainsi, plusieurs pays ont mis en pratique de réels systèmes utilisant la télémédecine pour l'assistance médicale prédictive. Nous indiquons que d'autres facteurs importants ont enrichi l'utilisation de cette évolution, il s'agit de la nouvelle génération de réseaux sans fil à savoir la 5G associée à l'intelligence artificielle [74].

Table. I.8. Tableau analytique de quelques articles de recherche sur l'exploitation du Data Mining en télémédecine

Références	Objective	Matériels et Méthodes
Mihaela Gheorghe et al. [75]	Consolider les données de divers systèmes (externes; internes) dans des BDD (base de données) distribués dans un entrepôt (datawarehouse)	ETC (extraction, transformation et chargement)
L. Lamberta,1, J. et al [76]	Ce projet SMART-EEG vise à modéliser une approche technique et pratique innovante utilisant les TIC pour l'interprétation des examens EEG via la télémédecine	Techniques de datamining
Michael Albert et al. [77]	L'objectif est de fournir à un expert à distance la capacité d'analyser les données des machines industrielles pour le diagnostic des pannes ou l'optimisation des processus de production. Ils ont décrit les modifications que nous avons apportées au système basé sur des règles pour pouvoir accomplir des tâches données pour l'analyse des données. Ilsonttraitéchaquetâche avec des donnéesenregistrées	Algorithmslogiciel de datamining
AbuKhousa E. [78]	Prévision des maladies en temps réel pour augmenter la capacité de réduction des risques des soins de santé primaires.	Techniques de datamining et IoT
Alessandro Massaro et al. [79]	gérer les activités de soins à domicile, la planification des visites et le service d'assistance, le diagnostic prédictif.	MLP-ANN
MihaMlakar et al. [80]	Le suivi des patients insuffisants cardiaques par concept objectif un modèle prédictif basé sur leurs sensations connues grâce à la mesure de leur : tension artérielle, température et humidité corporelle.	Forêt aléatoire (RF), Naïve Bayes (NB), arbre de décision (DT) et optimisation séquentielle minimale (SMO)
ALELYANI et al. [81]	Ce travail de recherche vise à augmenter la longévité et à la faire se situer entre 75 et 80 en arabe saoudien. Ce projet s'appelle " Saleem " (sain). C'est une plateforme qui s'intéresse aux diabétiques sous deux formes, soit pour suivre ceux qui le sont déjà, soit pour prévenir ceux qui risquent d'être diabétiques.	IoT; Smartphone; Datamining.
T. K. Anusuya1 et al.[82]	Fragmenter les bases de données médicales pour une utilisation à distance (Médecin-Patient) pour minimiser le coût.	1-Petits fragments disjoints. 2-regroupement de sites. 3-Allocation de fragments de données aux sites de cluster.
Salvatore itabile et al. [83]	Ce projet comporte trois axes : axe Big Data pour le suivi de santé à distance, axe pour le suivi lors d'un voyage et axe d'apprentissage automatique pour le suivi des activités de la vie quotidienne des personnes atteintes de démence.	Algorithme de clustering K-Means. L'algorithme de classification des forêts aléatoires ; IoT ; brouillard informatique
Ivanov, B., &Kralev, J.[84]	Les auteurs proposent une amélioration de la pompe à perfusion par une approche connectée à un patient qui lui assure un contrôle permanent et une circulation de son sang pour un traitement extracorporel afin d'aider les non spécialistes dans leur décision.	Techniques de datamining &capteurs
Gu, D. et al. [85]	Faites une étude de télémédecine pour les 20 dernières années sur la base de la base de données Web of science. Cette étude construit une structure de connaissances complète de la santé électronique dans le domaine de l'informatique de la santé. Celapeut aider de futursprojets de recherche.	Web des bases de données scientifiques
Lapointe L. et al. [86]	Ce travail montre les meilleures recherches (articles) qui aident les patients traumatisés en milieu rural ou dont le diagnostic est en retard en raison du manque d'experts utilisant la télémédecine et l'exploration de données.	1- Un formulaire d'extraction de données standardisé a été conçu en Microsoft Word. 2- Logiciel de gestion appelé « Zotero ».

Maria R. Bernstein et al. [87]	Conception d'un système de surveillance qui surveille la coagulation du sang chez les patients afin d'améliorer la gestion de la warfarine (traitement anticoagulant). Le principe de ce système est : 1-Consolidation des informations patient ; 2-Transmettre des données sur Internet ; 3-Et enregistrez ces données dans une base de données ; 4-Nettoyage de ces données pour obtenir des informations importantes sur les patients ; 5-Alerter le médecin en cas d'urgence et l'aider à prendre une décision appropriée et de qualité.	Techniques de datamining & Internet
ANGELUCCI, Alessandra et al. [88]	Cette recherche présente les meilleures technologies disponibles dans le domaine de la télésurveillance des patients respiratoires chroniques ainsi que la confidentialité des données et les bénéfices économiques et sanitaires.	Capteurs; techniques de Datamining et une excellente connexion ou 5G.
Thomas M. Judd et al. [89]	L'objectif de ce travail est d'intégrer un ensemble de services extra et intra hospitaliers dans un système médical de grande capacité pour la prise en charge qualitative d'un patient. Cette opération s'articule autour de deux domaines que sont le génie clinique (GC) et les technologies de l'information en santé (TIS). Techniquement, le travail consiste à analyser une grande masse de données concernant la génomique pour prédire les dysfonctionnements et l'amélioration de la qualité des soins. En conséquence, les meilleures pratiques actuelles du CEHTM et du CE-IT (ou TIC) ont été mises en évidence lors du récent troisième Forum mondial de l'OMS sur les dispositifs médicaux.	techniques de Bigdata
Clemens Scott Kruse[90]	L'objectif dans ce travail le rôle de la télémédecine consiste à suivre en temps réel les patients dépendants à l'alcool et aux drogues en analysant leurs données et en cas de dépression ou de complication mentale les assistants médicaux agissent en urgence.	L'Internet; L'exploration de données médicales et des articles ont été collectés à l'aide de mots clés tels que (télémédecine ; abus d'alcool et de drogues).
Rashid Bashshur et al. [91]	Une proportion importante des consultations externes dans divers contextes peut être gérée cliniquement efficacement à distance, c'est-à-dire que les patients atteints de maladies non urgentes peuvent être orientés vers un service de télémédecine sans compromettre leur santé ou la qualité des soins.	Internet; Smartphone; Medical datamining
Young-Rock Hong et al. [92]	Ils ont calculé la population américaine intéressée par les services de télésanté et le nombre d'hôpitaux qui ont approuvé les services de télésanté.	Google; Données médicales de l'American Hospital Association Annual 2018
Elie Massaad et al. [93]	Représenter et décomposer le volume, le contenu et la dispersion géospatiale des tweets liés à la télésanté pendant la pandémie de Covid-19.	l'exploration de texte sur Twitter ; traitement du langage naturel (TAL) ; méthodes d'apprentissage non supervisées (Clustering); Service d'informatique en nuage de Google
Katharina Boehm et al. [94]	a protégé les patients en urologie contre le COVID-19. Et offrir une continuité des soins sans contact par télésanté.	Internet; smartphone; data médical.

D'après le tableau I.8, les auteurs de [75] décrivent l'intégration des méthodes d'exploration de données dans les systèmes de télémédecine. Ils présentent également une vue d'ensemble pour comprendre et améliorer la solution développée à l'aide des techniques de gestion des processus métier. Dans les travaux de recherche présentés dans [76], les auteurs proposent une approche de la télétransmission en électroencéphalographie appelée SMART-EEG.

Il s'agit d'une solution innovante de transmission à distance d'EEG qui optimiserait l'utilisation des ressources humaines et des compétences dans son champ d'application. Il met à disposition des patients éloignés des médecins compétents en déficit démographique, facilite l'accès à des soins de qualité pour tous et limite les transports et hospitalisations inutiles et coûteux. Dans les travaux [77], les auteurs définissent « une extension d'une approche de télémédecine, conçue pour analyser les données des machines industrielles, et fournir un expert à distance pour le diagnostique des pannes ou pour l'optimisation des processus de production.

Une nouvelle vision a fait l'objet du travail décrit dans [78]. Il s'agit d'un environnement de soins primaires intelligent (IPCE) présenté comme un écosystème de santé composé de différents objets interagissant les uns avec les autres, par exemple des capteurs, des dispositifs, des systèmes intégrés pour des soins primaires intelligents. Les travaux de recherche décrits dans [79] ont présenté une plateforme innovante de gestion des ressources (RMP), en employant un système d'aide à la décision (DSS) basé sur un perceptron multicouche (MLP) de réseau neuronal artificiel (ANN). Cette plateforme est capable de transférer efficacement les analyses des patients de leur domicile vers la salle de contrôle d'une clinique, avec un coût réduit et une qualité de services supérieure. Pour une prise en charge rapide dans le domaine cardiaque, dans [80], les auteurs ont donné un cahier des charges pour une surveillance télécommandée du système d'extraction de données physiologiques pour les patients souffrant d'insuffisance cardiaque congestive. Ce travail a été réalisé sur deux groupes d'hommes et de femmes souffrant d'insuffisance cardiaque de deux pays, l'Italie et le Royaume-Uni. L'âge des patients est compris entre 9 et 62 ans, 15 classes NYHA II et 9 classes NYHA III, 10 ischémies, étiologie, 6 dilatées, 2 valves et 6 polymorphes.

Dans l'article [81], les auteurs proposent une plateforme de télémédecine pour le contrôle et la gestion du diabète. Cette plateforme permet aux patients diabétiques d'enregistrer et de suivre leur maladie et leurs signes vitaux. De plus, cette plate-forme fournit aux patients des commentaires et des notifications afin de les aider à contrôler et à atténuer les complications du diabète. De plus, le médecin du patient dispose d'un tableau de bord de suivi et de communication pour suivre la situation des patients et communiquer avec eux si nécessaire. Saleem est un système intelligent conçu pour illustrer le comportement quotidien des patients diabétiques. Dans le travail [82], les auteurs décrivent une solution Web comme un service de télémédecine basé sur le datamining. Les auteurs définissent un système d'archivage et de communication appelé PACS, il est utilisé pour le processus de télémédecine, lorsque les antécédents du patient, les images médicales ou toute autre information connexe sont nécessaires. [83]. Dans l'article [84], les auteurs proposent une solution basée sur le concept de Bigdata. Il s'agit d'un système de surveillance de la santé à distance, exploitable dans un cloud public, privé ou hybride.

Ils expliquent également une approche pour soutenir la surveillance distribuée de la santé humaine pendant un voyage. Une autre partie du travail a porté sur l'utilisation de l'apprentissage automatique pour surveiller les activités de la vie quotidienne des personnes atteintes de démence. Les auteurs dans [85], décrivent le processus de transfert de données entre les différentes équipes hospitalières pour assurer un traitement efficace et prendre la bonne décision. Dans l'article [86], les auteurs ont proposé un aperçu des publications et contributions scientifiques concernant l'application des TIC, de l'IoT et du big data dans le secteur des soins. Ils ont travaillé sur 3085 articles publiés de 1992 à 2017.

Un autre travail dans [87] a été réalisé comme une étude qui visait à évaluer l'impact de la télémédecine sur le diagnostic, la prise en charge clinique et les résultats chez les patients traumatisés dans les zones rurales. Dans les travaux développés dans [88], les auteurs s'intéressent aux services de gestion de l'anticoagulation avec le modèle de base de l'analyse des données. Ils ont constaté qu'une prise en charge optimale par rapport aux soins médicaux traditionnels devrait éviter 7 infarctus du myocarde, accidents vasculaires cérébraux, saignements majeurs ou décès pour 100 patients / an. Dans [89], les auteurs insistent sur des systèmes de télésurveillance pour aider les patients respiratoires chroniques.

Généralement, ces systèmes suivent l'architecture à deux bonds. Dans lequel, les données sont transmises des capteurs à une passerelle avec des technologies de liaison capteur-gestionnaire, puis elles sont transmises de la passerelle à la gestion des données par des technologies de liaison cellulaire. Les appareils de spirométrie portables peuvent être considérés comme le modèle le mieux adapté pour les applications de surveillance à distance.

Nous concluons cette section sur l'analyse des données et son lien étroit avec la télémédecine en présentant quelques travaux récents qui expliquent l'importance et l'apport de la télémédecine dans le monde à travers des analyses de données.

A cet effet, dans les travaux [90], les auteurs donnent une description concernant l'ingénierie clinique (CE) qui est plus que jamais étroitement associée aux technologies de l'information de santé (TIS). Dans leur article, ils décrivent une série de mises en œuvre et de stratégies pour leur utilisation efficace pour fournir des services de soins de haute qualité et des résultats cliniques supérieurs.

Les meilleures pratiques actuelles en matière de CEHTM et de CE-IT (ou TIC : Technologies d'Information et de Communication) ont été mises en avant lors du récent troisième Forum mondial de l'OMS (Organisation Mondial de la Santé) sur les dispositifs médicaux.

Dans les travaux de recherche présentés dans [91], l'accent est mis sur l'évaluation de l'efficacité et de la qualité de la télémédecine pour surveiller la toxicomanie, l'abus d'alcool et la réadaptation dans trois régions des États-Unis d'Amérique.

Les résultats obtenus indiquent que l'utilisation de la télémédecine a aidé 59% des alcooliques à éliminer la consommation d'alcool, réduit la dépression de 11%, augmenté la satisfaction des patients de 9%, augmenté l'accessibilité de 9%, amélioré la qualité de vie de 6% et diminué les coûts de 3%. Les interventions liées à la santé mobile étaient de 50%, la e-santé à 27%, le téléphone à 14% et la vidéo bidirectionnelle à 9%. Des études européennes ont été menées à 36% et australiennes à 5%. Les trois derniers travauxsont liés au COVID-19 où le datamining et la télémédecine sont très utilisés et ils ont joué un rôle important pour aider et gérer la situation de la pandémie. Ces travaux nous permettent d'observer qu'au fur et à mesure que le nombre de cas de COVID-19 augmente, l'intérêt de la population pour la télésanté augmente également. S'il n'y a actuellement aucun remède médical à la pandémie COVID-19, la seule stratégie possible pour réduire l'infection en plaçant des mesures et des protocoles médicaux : éloignement social, quarantaine, etc. Dans ce cas, la télésanté est très importante pour aider les patients, et fournir les différents besoins psychologiques et physiologiques, cependant, téléphone, e-mail, visioconférence…). De plus, la télésanté réduit le contact de personne à pasteur, elle est donc très appropriée pour fournir des services médicaux aux patients Covid-19 et aux patients susceptibles. Les résultats des études suggèrent que la télémédecine améliorerait le diagnostic, simplifierait le processus de transfert des patients et raccourcirait le séjour. La télémédecine a eu un effet minime sur la mortalité et les complications chez les patients traumatisés dans les zones rurales

I.6 Inconvénients et avantages du Data Mining en médecine

Après cette étude, nous avons extrait quelques inconvénients et avantages du datamining [95] dans le secteur de la santé, que nous présentons ci-dessous.

I.6.1. Désavantages

Parfois, l'exploration de données fournie est impure et entraîne donc de mauvaises décisions. Parmis ses limites nous citons :

- la variété de types de données médicales associées à un très grand volume ;
- la difficulté d'interprétation des résultats par des experts médicaux ;
- la mauvaise catégorisation mathématique ;
- Le stockage des données des patients soulève un problème de confidentialité d'un point de vue juridique et social.
- Le partage centralisé des données pose le problème du partage de ces données avec différents utilisateurs.
- Certaines données sont imprécises, incertaines et même incomplètes.
- Les approches d'exploration de données sont parfois inefficaces, ce qui conduit parfois à des décisions risquées.

I.6.2. Avantages

L'exploration de données présente certains aspects positifs ; nous pouvons les résumer comme suit :

- L'efficacité du datamining dans le processus de diagnostic et de pronostic des maladies.
- Déterminez l'interaction entre plusieurs maladies.
- Aide les experts à définir les traitements appropriés et le suivi le plus adapté.
- Fournir des services de santé meilleurs et plus adéquats.
- L'exploitation rapide d'un grand volume de données.
- Minimisation le coût du traitement.
- Génération et analyse plus rapides des rapports de synthèse.
- Les informations prédictives peuvent être extraites de grandes bases de données,
- La centralisation des données qui permet de varier les traitements.
- Multiplicité des applications d'analyse de données.

I.7. Conclusion

Dans ce chapitre nous avons dressé un état de l'art sur le domaine du datamining et son importance en matière d'analyse de données pour des secteurs manipulant un grand volume de données. Nous avons par la suite mis l'accent sur l'importance du datamining dans le domaine medical et plus exactement dans la télémedicine.

A cet effet, nous avons d'abord explicité le domaine de datamining ainsi que les approches les plus utilisées spécialement dans la médecine suivie d'une synthèse des travaux existants.

Par la suite, nous avonsdressés une présentation de la télémédecine ainsi que les travaux exploitant le datatmining pour répondre aux décideurs du domaine médicale en matière d'extraction de connaissance vitale pour l'aide à la décision prédictive.

Dans le chapitre suivant, nous décriveront les fondements théoriques de la métaheuristiqueSOS (SymbioticOrganismsSearch) pour une utilisation comme approche d'analyse de données dans le domaine médical à travers l'utilisation du datamining.

Chapitre II

Première contribution :
Expérimentation et évaluation de SOS
(Symbiotic Organisms Search)

II.1 Introduction

Manipuler des données très volumineuses, afin de prendre la meilleure décision, n'est possible que grâce à une extraction de connaissances. L'exploration de données est devenue un processus largement utilisé dans l'analyse de données pour extraire les connaissances les plus importantes pour la prise de décision prédictive. L'un des types importants d'exploration de données est le mécanisme de clustering ; son but est de diviser les données très volumineuses en un ensemble de clusters, le nombre de paramètres est très élevé et le problème de clustering est plus difficile [1][2].

Les métaheuristiques ont été largement utilisées dans le regroupement ; ils peuvent apporter des solutions satisfaisantes à des problèmes complexes. L'objectif principal de ce chapitre est de proposer un nouvel algorithme de clustering basé sur une technique métaheuristique appelée recherche d'organismes symbiotiques (Symbiotic Organisms Search : SOS), il s'inspire d'un processus biologique et simule l'interaction symbiotique entre les organismes d'une même population [10].

Plusieurs travaux de rechèrche ont exploité SOS dans differents domaines qui confirment son efficacité et sa robustesse pour résoudre les problèmes de clustering de données. De plus, SOS n'a pas de paramètres de contrôle qui nécessitent un ajustement ; ce qui rend l'utilisation de SOS pour résoudre le problème de clustering plus facile sur différents types de bases de données [11].

La méthode SOS est utilisée pour trouver les centres optimaux d'un certain nombre de clusters, en tant que technique d'exploration de données supervisée. Les résultats expérimentaux ont été réalisés en deux phases. Premièrement, la technique SOS est étalonné avec six fonctions de test bien connues. Deuxièmement, différents ensembles de données médicales ont été utilisés pour tester notre méthode de clustering proposée basée sur SOS ayant pour objectif de montrer sa crédibilité du traitement.

II.2 Spécification de l'algorithme de recherche d'organismes symbiotiques (SOS)

SOS est un algorithme métaheuristique recemment développé par Cheng en 2014. Il est bio-inspiré par la dépendance basée sur l'association durable entre deux ou plusieurs organismes de l'écosystème et profitable à chacun d'eux, tous deux bénéficiant de cette association.SOS est comme la plupart des algorithmes métaheuristiques basés sur la population, il présente les caractéristiques suivantes [10] :

- Des solutions candidates qui constituent une population d'organismes dispersés dans l'espace de recherche ;
- Des solutions candidates qui sont les éléments clés des opérateurs pour l'orientation de la recherche ;

• Une stratégie sélective est exploitée pour obtenir les meilleures solutions ;

• Un mécanisme de validation des caractéristiques intrinsèquement importantes, à savoir la taille de la population et le nombre maximum d'évaluations.

Cependant, contrairement à la plupart des algorithmes métaheuristiques qui ont des paramètres de contrôle (par exemple, AG a le taux de croisement et de mutation ; DE a également des paramètres de croisement et de mutation ; le PSO est caractérisé par des aspects spécifiques tels que le poids d'inertie, les facteurs cognitifs et les facteurs sociaux), [10] [11]. SOS ne nécessite aucun paramètre spécifique. Ceci est considéré comme un avantage par rapport aux algorithmes concurrents [10]. Le réglage imprécis des paramètres spécifiques à l'algorithme peut augmenter le temps de calcul et aboutir à des solutions optimales locales [10]. Le processus SOS comprend trois phases : mutualisme, commensalisme et parasitisme.

II.2.1. Phase de mutualisme

Dans cette étape, les organismes s'entraident pour assurer une existence dans l'écosystème. Le processus de survie est basé sur un changement d'état de chacun des deux organismes candidats X_i et X_j exprimé par les équations suivantes [10] :

$$X_{i\_new} = X_i + \lambda\ (X_{best} - \text{Vecteur\_Mutuel} \times BF1) \quad \text{(Eq.13)}$$

$$X_{j\_new} = X_j + \beta\ (X_{best} - \text{Vecteur\_Mutuel} \times BF2) \quad \text{(Eq.14)}$$

Où :

$$\text{Vecteur\_Mutuel} = (X_i + X_j) / 2 \quad \text{(Eq.15)}$$

La relation symbiotique, également appelée caractéristique, entre les organismes X_i et X_j, est représentée par un vecteur appelé "Vecteur_Mutuel ". BF1 et BF2 sont les facteurs de gain, ils sont choisis au hasard comme 1 ou 2. Chaque facteur est le niveau de gain partiel ou complet de chaque organisme. La meilleure solution d'adaptation de l'organisme est représentée par X_{best}. λ et β sont des valeurs aléatoires choisies dans l'intervalle [0,1].

II.2.2 Phase de commensalisme

Dans cette phase, l'un des deux organismes obtient un gain mais l'autre va sans rien gagner ni rien perdre. L'organisme X_j représente celui qui ne gagne ni ne perd rien de la relation, et le nouvel organisme X_i est calculé selon la symbiose du commensalisme entre les organismes X_i et X_j. Son expression est donnée comme suit [10]:

$$X_{i\_new} = X_i - \alpha\ (X_{best} - X_j) \quad \text{(Eq.16)}$$

Où : α représente un nombre aléatoire dans [-1,1] et X_{best} est le degré le plus élevé de l'organisme d'adaptation.

II.2.3 Phase de parasitisme

Le parasitisme est un comportement dans lequel un organisme tire profit d'une relation mais nuit activement à l'autre organisme. Un parasite artificiel appelé "Vecteur_Parasite" est créé dans l'espace de recherche en utilisant la duplication de l'organisme X_i puis modifié les dimensions concernées aléatoirement par un nombre aléatoire.

Ce Vecteur_Parasitetente ($X_{parasite}$) de remplacer un autre organisme X_j dans l'écosystème. Pour ce but, si Vecteur_Parasiteest meilleur, il tuera l'organisme X_j et prendra sa position ; sinon, X_j sera immunisé contre le parasite et ne pourra plus exister dans cet écosystème [10].

$$X_{parasite} = \text{rand}(0,1) \times (UB - LB) + LB \quad \text{(Eq.17)}$$

Où UB, LB sont respectivement la borne supérieure et la borne inférieure du problème d'optimisation considéré.

Les trois phases sont formalisées par l'algorithme suivant :

```
Initialisation (ecosystème, fonction fitness, itération maximun)
Pour contrer=1 : itération maximum
    Pour contrer=1 : taille (ecosystème)
        Phase Mutualisme
        Phase Commensalisme
        Phase Parasitisme
        Identifier de meilleurs organismes
    Fin Pour
Fin Pour
```

Figure. II.8. Pseudo-code de l'algorithme SOS

II.3. Modélisation du processus de Data Mining

De nombreuses informations circulent dans l'industrie de l'information en général, et dans tout secteur médical, économique ou administratif. Ces données n'ont aucune valeur cognitive ou économique tant qu'elles ne sont pas converties en informations utiles. Il est intéressant de voir la nécessité d'analyser ce grand volume de données et d'extraire des informations utiles et exploitables [2]. L'exploration de données est le processus de récupération d'informations à partir d'énormes magasins de données. En d'autres termes, l'exploration de données est également définie pour extraire des connaissances [2].

Les données sont généralement présentées dans une grande matrice (X) qui contient des informations dans un champ spécifique, où le nombre de ses lignes représente les objets ou les paramètres (n) et le nombre de ses colonnes est les fonctionnalités (m).

$$X = \begin{Bmatrix} x_{11}\ x_{12}.....x_{1m} \\ x_{21}.....x_{2m} \\ ..... \\ ..... \\ ..... \\ x_{n1}.....x_{nm} \end{Bmatrix}$$

Cette matrice doit être classée en divers groupes d'objets similaires, et chaque groupe est différent de l'autre. Il existe plusieurs méthodes pour le faire, mais nous nous intéressons aux techniques de clustering.

II.4. Clustering (Regroupement)

L'algorithme de clustering traite un groupe de jeu de données (supposé être défini comme une matrice X). Au cours de l'analyse complète, il tente de localiser l'ensemble de données en groupes via un ensemble de k classes, telles que $C = \{c_1, c_2,\ldots, c_k\}$, k représente le nombre de clusters. Généralement, la technique de clustering dépend de la relation entre des ensembles de données similaires, son but est de trouver une forte similitude entre les membres d'un même cluster et une faible similitude entre les membres de différents clusters [31].

II.5. Proposition de SOS pour le clustering

Dans notre proposition, nous avons utilisé SOS dans un processus de clustering [31]. L'idée de base repose sur la modélisation du principe de fonctionnement du clustering supervisé. C'est une technique de data mining par la méthode SOS qui est associée à une fonction de fitness définie dans la siction **II.6**. L'algorithme proposé se compose de quatre étapes principales.

Dans la première étape, nous sélectionnons le jeu de données correspondant X qui contient n organismes $x_1, x_2,\ldots ..x_n$. Chacun a m fonctionnalités ($x_i = \{x_{i1}, x_{i2},\ldots .., x_{im}\}$), nous définissons le nombre de classes k, et les centres de clustering $C = \{c_1, c_2,\ldots .., c_k\}$. Dans un deuxième temps, l'approche encode les organismes de l'écosystème et génère la solution de regroupement finale basée sur la prise en compte de chaque organisme de l'écosystème comme une solution de regroupement complète.

Nous définissons la longueur de chaque organisme (Dim) par l'expression suivante : $Dim = k \times m$. La figure II.9 montre codage de l'organisme. La troisième étape consiste en un processus de classification basé sur des organismes caractéristiques. Il développe une technique de regroupement utilisant le principe de la distance minimale grâce à la fonction de fitness (objective) qui calcule la distance entre chaque organisme et le centre de regroupement (C).

Dans la dernière étape, le processus regroupe les paramatres qui sont en étroite relation les unes avec les autres, bien que l'algorithme de clustering essaie de minimiser la fonction de fitness (*ff*) pour obtenir les meilleurs résultats de clustering. Ce cycle est répété plusieurs fois jusqu'à ce que le meilleur classement soit obtenu [33].

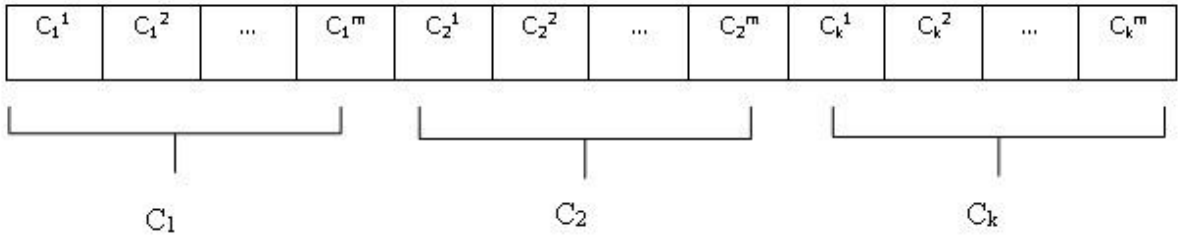


Figure. II.9. Codage de l'organisme

Les étapes de la méthode de clustering SOS proposée sont spécifiées respectivement dans l'architecture et l'organigramme suivants [33] :

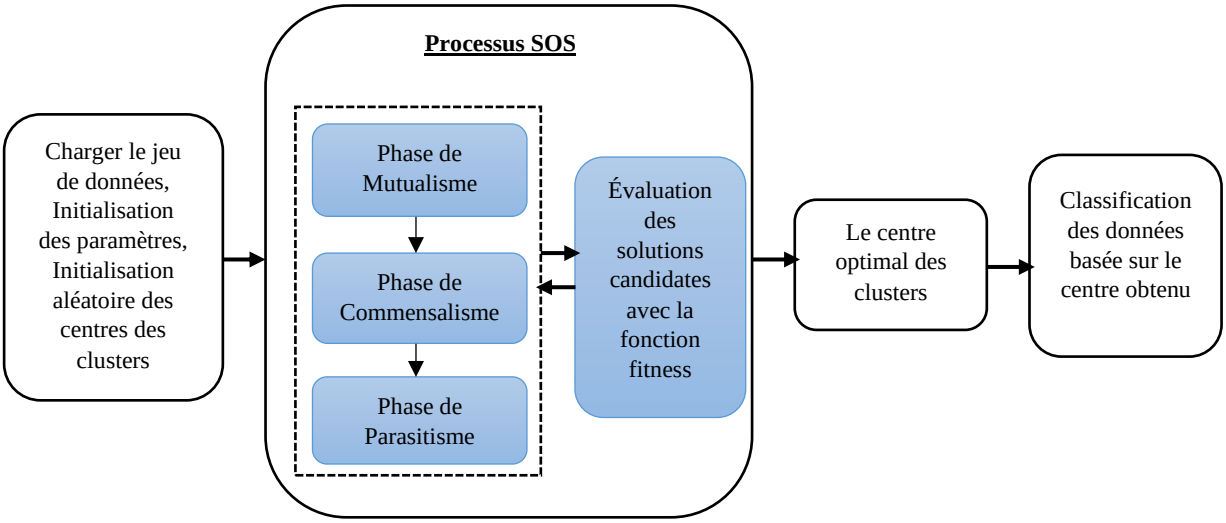


Figure. II.10. Architecture de l'algorithme de clustering SOS

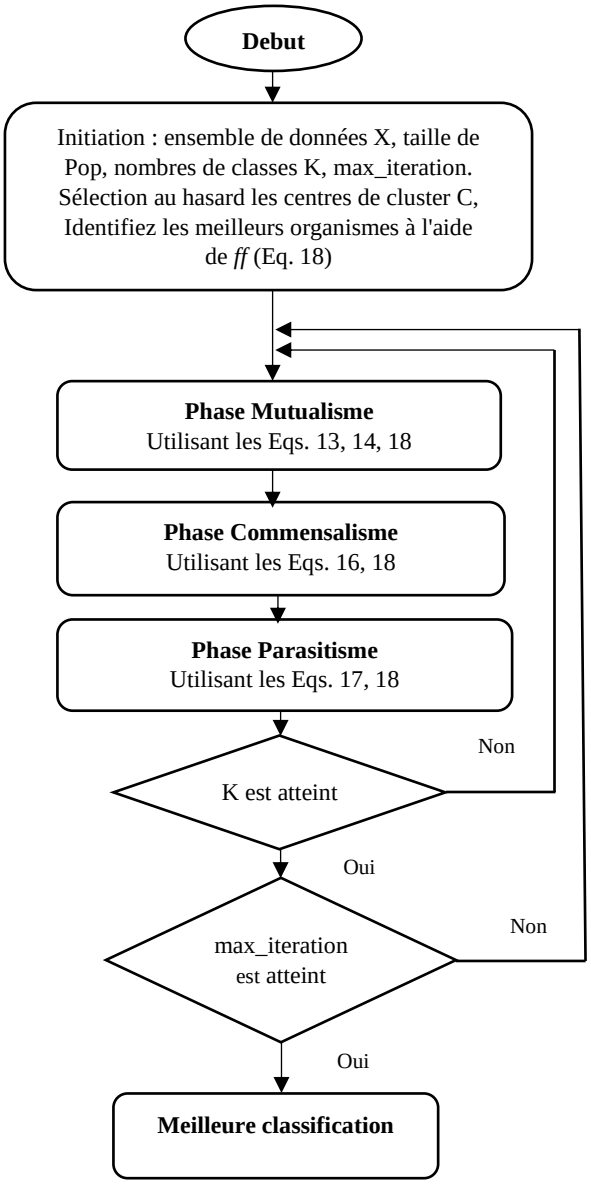


Figure. II.11. Organigramme de l'algorithme de clustering SOS

II.6. La fonction de fitness (objective)

Dans cette section, nous décrivons une méthode de clustering supervisé. Dans un premier temps, l'algorithme génère aléatoirement les centres des classes, puis nous appliquons la fonction fitness qui est utilisée dans les trois phases. La qualité des résultats obtenus dépend directement de la spécification formelle de la fonction objective.

L'étude des fonctions existantes nous a donné l'inspiration d'une fonction de fitness ayant comme paramètre fondamental la distance minimale entre les paramatres et les centres des classes (C) proposée aléatoirement au départ.

La formule de cette fonction de fitness est la suivante [33] :

$$ff = \sum_{i=1}^{n} \sum_{j=1}^{k} (\text{dmin}\ (X_i, c_j) \qquad \text{(Eq.18)}$$

Où:

dmin : distance minimale entre chaque instance et les centres de classe.

X : matrice de jeu de données.

K : nombre de classes.

C : les centres de classe, tels que C = $\{c_1, c_2, \ldots .., c_k\}$.

n : nombre des instances ou population de l'ensemble de données.

II.7. Expérimentation et discussion sur les résultats

Dans cette section, l'algorithme SOS est testé avec six fonctions de référence bien connues [97], ses résultats sont comparés à ceux obtenus à l'aide d'autres algorithmes et qui sont :ABC (ArtificialBee Colony) [98], CA (Cultural Algorithms) [99], GWO (Grey Wolf Optimizer), PSO (Particle Swarm Optimizers) [100], CDWPSO (Chaotic Dynamic Weight Particle Swarm Optimization) [101] et BOA (Butterfly Optimization Algorithm) [102].

Par la suite, le clustering basé sur la méthode SOS est évaluée à l'aide de quatre bases de données médicales réelles et publiques pour les classifications.

Les fonctions de test ont été sélectionnées de trois catégories différentes [97]:

• Catégorie 1 : fonctions unimodales à haute dimension : la fonction ‘Schwelef 2.21’ (F01) et la fonction ‘’Schwelef 2.22’’ (F02), toutes deux de dimension 30.

$$F_{01} = \sum_{i=1}^{n} |x_i| + \prod_{i=1}^{n} |x_i|$$

$$F_{02} = \ max_i\{|x_i|, 1 \le i \le D\}$$

• Catégorie 2 : fonctions multimodales à haute dimension : fonction ‘Rastrigin’ (F03) et fonction ‘Ackley’ (F04) toutes deux de dimension 30.

$$F_{03} = \ \sum_{i=1}^{D} \left[x_i^2 - 10\ cos(2\pi x_i) + 10\right]$$

$$F_{04} = \ -20\ exp\left(-0.2\sqrt{\frac{1}{D}\sum_{i=1}^{D} x_i^2}\right) - exp\left(\frac{1}{D}\sum_{i=1}^{D} cos(2\pi x_i)\right) + 20 + e$$

• Catégorie 3 : Fonctions multimodales de faible dimension : fonction ‘Six Hum’ ou Quartic avec bruit (F05) de dimension 4 et fonction ‘Goldstein’ (F06) de dimension 2.

$$F_{05} = \sum_{i=1}^{11}\left[a_i - \frac{x_1\left(b_i^2 + b_i x_i\right)}{b_i^2 + b_1 x_3 + x_4}\right]^2$$

$$F_{06} = [1 + (x_1 + x_2 + 1)^2(19 - 14x_1 + 3x_1^2 - 14x_2 + 6x_1x_2 + 3x_2^2)] \times [30 + (2x_1 - 3x_2)^2(18 - 32x_1 + 12x_1^2 + 48x_2 - 36x_1x_2 + 27x_2^2)]$$

Les ensembles de données de classification utilisés dans la deuxième expérimentation son t: le cancer du sein (BC), les maladies cardiaques, les troubles hépatiques et le diabète indien Pima obtenus à partir du référentiel d'apprentissage automatique de l'Université de Californie à Irvine (UCI) [103].

Pour une excellente estimation et appréciation des performances de la méthode de clustering SOS, l'algorithme est comparé aux algorithmes GA [104], PSO [100] et DE (Differential Evolution) [99]. Pour évaluer nos résultats, les mesures de comparaisons utilisées sont : la matrice de confusion, les taux d'exactitude de la classification et les courbes ROC.

Dans le tableau II.9, les différentes fonctions utilisées pour le processus de test sont répertoriées. Cette liste de fonctions nous a permis de définir les dimensions, le nombre maximum d'itérations et l'intervalle de solutions.

Ces fonctions ont été choisies car elles sont complexes et l'obtention de leurs solutions optimales globales est difficile en raison de l'existence de plusieurs optimums locaux.

Table. II.9. Caractéristiques fonctionnelles des fonctions de teste

Catégorie	**Nom des fonctions de test**	**Dimension**	**Intervalle**	**Itération maximale**	**Optimum**
Unimodal functions with high dimensions	F1: Schwelef 2.21	30	[-10,10]	950	0
	F2 : Schwelef 2.22	30	[-100,100]	1000	0
Multimodal functions with high dimensions	F3 : Rastrigin	30	[-5.12,5.12]	40	0
	F4 : Ackley	30	[-32,32]	60	8.8818e-16
multimodal functions with low dimension	F5 : Six hum	4	[-5,5]	400	0.0003075
	F6 : Goldstein	2	[-5,5]	200	3

Des courbes de fonctions à trois dimensions sont présentées sur la figure II.12, ce type de courbes nous montre clairement le nombre d'optimum locaux pour chacune d'elle.

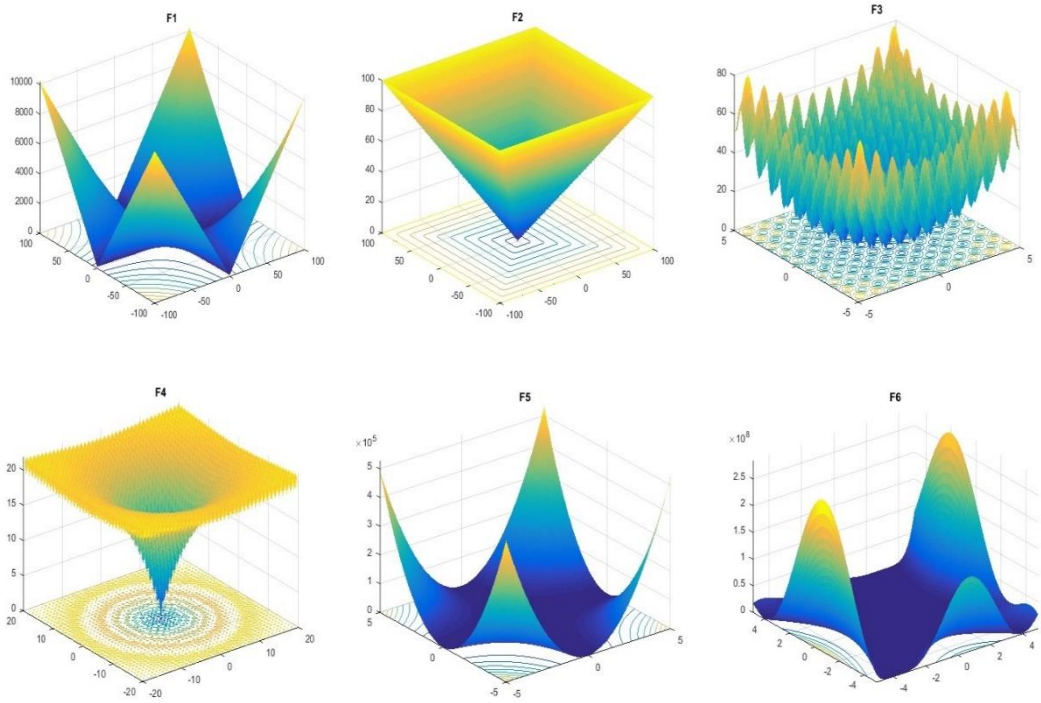


Figure. II.12. Courbes à 3D des fonctions utilisées

II.7.1. Etude comparative

Pour avoir une bonne appréciation de notre approche proposée, la méthode de recherche d'organismes symbiotiques (SOS) est comparée à d'autres méthodes : PSO, ABC, CA, GWO, CDWPSO et la nouvelle méthode BOA.

Les résultats ont été obtenus après 25 exécutions indépendantes et sont présentés dans le tableau II.10. Pour l'évaluation de chaque fonction, nous avons utilisé trois critères qui sont la moyenne (Moy), l'écart type (StD) et le classement de chaque méthode (Rang).

Table. II.10. Résultats de l'évaluation des méthodes

Fonction	Parametres	PSO	ABC	CA	GWO	CDW-PSO	BOA	SOS
F1	Moy	0.0352	0.1752	59.2268	8.039e-47	**0**	6.99 e−153	**0**
	StD	0.1624	0.0990	20.9776	8.345e-47	0	1.47 e−152	0
	Rang	5	6	7	4	**1**	3	**1**
F2	Moy	2.1999	12.3781	3.1395	1.384e-21	**0**	**0**	**0**
	StD	1.7124	1.6148	3.2499	2.856e-21	0	0	0
	Rang	5	7	6	4	**1**	**1**	**1**
	Moy	275.2287	308.3371	316.9494	47.6089	2.3177	2.883 e+1	**0**

F3	StD	17.2402	20.9358	18.6367	15.1017	8.2990	3.128 e−2	**0**
	Rang	5	6	7	4	2	3	**1**
F4	Moy	13.4428	16.4168	11.1855	0.0844	0.9060	1.7183	**8.881 e-16**
	StD	0.7387	0.7822	0.8800	0.0292	1.6656	2.688 e−2	**0**
	Rang	6	7	5	2	3	4	**1**
F5	Moy	0.0036	1.0010	0.0044	0.0019	5.96e-4	3.8917 e−5	**3.441 e-04**
	StD	0.0075	3.396e-5	0.0122	0.0056	6.14e-4	2.9003 e−5	**1.831 e-04**
	Rang	6	4	7	5	3	2	**1**
F6	Moy	3.000	3.001	3.000	3.0000	8.1731	**3**	**3.0000**
	StD	1.896 e-12	1.8513 e-4	1.3628 e-15	2.1140e-5	5.3154	**0**	**0.536 e-16**
	Rang	4	6	3	5	7	**1**	**2**
Rang intermédiaire		5.166	6	5.833	4	2.833	2.333	1.166
Classement général		**5**	**7**	**6**	**4**	**3**	**2**	**1**

Ce classement a été obtenu sur la base de la meilleure moyenne des valeurs des fonctions objectives et des meilleurs écarts-types par rapport à la valeur optimale des trois catégories.

- Le tableau II.10 montre que SOS a trouvé l'optimum global exact pour les fonctions F1 à F4, mais pour les deux fonctions F5 et F6 il l'a trouvé avec une valeur d'écart type assez petite (nous le considérons comme une erreur).
- Une valeur moyenne proche de l'optimum prouve l'efficacité de la méthode, nous notons aussi une valeur plus petite de l'écart type confirme sa stabilité.
- Cela signifie que SOS converge parfois vers l'optimum local.
- Généralement, la méthode SOS donne pour F1 à F6 des solutions optimales par rapport aux autres méthodes, elle prend le premier rang, uniquement pour F6 où la nouvelle méthode BOA est classée en premier.

II.7.1.1 Courbe de forme physique

Pour montrer plus de performances de l'algorithme SOS, nous présentons les courbes des fonctions de fitness. La figure II.13 montre les courbes de convergence obtenues par différentes méthodes. Nous n'avons sélectionné que deux fonctions (F3 et F6) qui contiennent plusieurs optimums locaux.

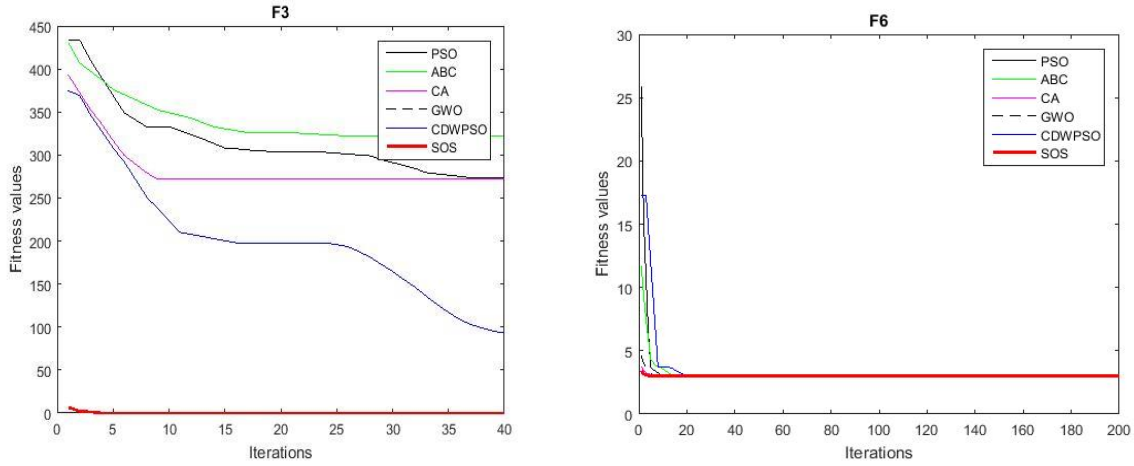


Figure II.13. Courbes de fitness pour 2 fonctions comparées par 6 méthodes

Les courbes illustrées sur la figure II.13 représentent une comparaison des résultats de fitness des deux fonctions selon SOS, PSO, ABC, CA, GWO et CDW-PSO. On remarque que SOS converge très rapidement et qui atteint l'optimum global mais les autres méthodes notamment pour la fonction F3 convergent vers un optimum local.

II.7.1.2 Analyse des variations (ANOVA)

Pour expliquer notre évaluation de la méthode proposée par rapport aux autres algorithmes de comparaison, nous avons utilisé sur la figure II.14 le 'test ANOVA' qui est une représentation schématique de la distribution d'une variable. Le test Anova inventé en 1977 par John Tukey, également appelé Tukey box ou box-and-whisker plot est utilisé pour représenter le schéma essentiel d'une série statistique quantitative [105].

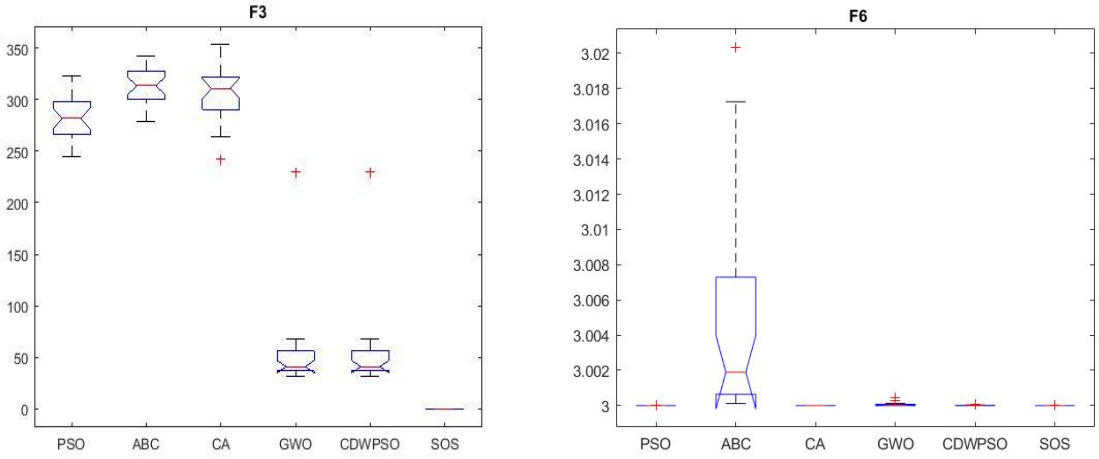


Figure II.14. Test Anova de 2 fonctions pour les 6 algorithmes utilisés**.**

Les formes de la figure II.14 obtenues par ce test, qui sont les cases, représentent les variations des solutions optimales entre les valeurs de fitness minimum et maximum, respectivement en bas et en haut de la case, centrées sur la médiane.

Ce centre représenté par la ligne rouge de la case est la valeur moyenne des 25 exécutions des différents algorithmes, qui est la solution optimale de la fonction sujet à tester. Ainsi notre constat est :

- On remarque que SOS est plus stable que les autres algorithmes ;
- À partir de tous les résultats ci-dessus, il est clair que la méthode SOS a une meilleure précision de recherche que les cinq autres méthodes ; ils indiquent également que l'algorithme SOS a une vitesse de convergence plus rapide ;
- Cela prouve que SOS a une forte robustesse et une excellente stabilité assurant une convergence claire par rapport aux autres méthodes ;

Par conséquent, nous concluons que l'algorithme SOS peut résoudre des problèmes de grande dimension.

II.7.2. Évaluation SOS pour l'exploration de données

Après avoir vérifié la robustesse de SOS, nous allons maintenant faire une évaluation de l'algorithme SOS pour le mécanisme de clustering automatique dans l'exploration de données.

Les expériences réalisées avec l'algorithme de clustering SOS visent à classer les bases de données de manière supervisée. Pour une évaluation qualitative, les mêmes expériences sont réalisées avec les algorithmes GA, DE et PSO sur les mêmes ensembles de données.

II.7.2.1 Bases de données utilisées pour l'exploration de données

Pour montrer que l'approche de clustering à base deSOS est une excellente technique pour l'exploration de données, nous avons utilisé les quatre mêmes ensembles de données cités précédemment et détaillés comme suit [103]:

- ***Base de données sur le cancer du sein (Breast Cancer : BC)***

Cette base de données a été préparée par William H. Walberg des hôpitaux de l'Université du Wisconsin, Madison.

Elle contient 683 instances et 11 attributs (ou caractéristiques) tels que: l'épaisseur de l'agrégat, l'uniformité de la taille des cellules, l'uniformité de l'adhérence marginale de la forme des cellules.Ils sont bien détaillés dans le tableau II.11.

- ***Base de données sur les maladies cardiaques (Heart Disease : HD)***

Cette base de données a été créée pour le traitement et l'analyse décisionnelle des images de tomographie par ordinateur à proton unique (SPECT).

Cette base de données contient 267 images, 80 instances et 22 fonctionnalités (voir le tableau II.12). Toutes les fonctionnalités sont en binaire (0 ou 1). Après le traitement, le diagnostic déclare si le patient est normal ou non.

- ***Base de données sur les troubles hépatiques (Liver Disorder : LD)***

La base de données sur les troubles hépatiques a été fournie par BUPA MedicalResearch Ltd pour les troubles hépatiques sous la forme d'une étiquette binaire. Elle contient des valeurs de 7 attributs mesurés pour 345 patients de sexe masculin. Les cinq premiers représentent les données de tests sanguins considérées comme sensibles aux troubles hépatiques dus à une consommation excessive d'alcool, le tableau II.13 montre les fonctionnalités de cette base.

- ***Base de données de Pima Indian Diabetes (PID)***

Cette base de données concerne les femmes enceintes âgées d'au moins 21 ans et d'origine indienne. Pima contient un nombre d'instances égal à 768 avec un nombre d'attributs égal à 8. Le tableau II.14 présent des informations détaillées sur les fonctinnalités de la base PID.

Tableau . II.11. Description des attributs de l'ensemble de données sur les maladies des patients de cancer du sein

Attribut	Description	Interval
Épaisseur de touffe	Évalue si les cellules sont mono ou multicouches.	0-10
Uniformité de la taille de la cellule	Évalue la cohérence de la taille des cellules de l'échantillon.	0-10
Uniformité de la forme de la cellule	Estime l'égalité des formes de cellules et identifie les variances marginales.	0-10
Adhérencemarginale	Quantifie la quantité de cellules à l'extérieur de l'épithélium qui ont tendance à se coller les unes aux autres.	0-10
Taille de cellule épithéliale unique.	Se rapporte à l'uniformité cellulaire, détermine si les cellules épithéliales sont considérablement agrandies.	0-10
Noyaux nus.	Présence et taille des noyaux.	0-10
chromatine fade.	Évalue la « texture » uniforme du noyau dans une plage allant de fine à grossière.	0-10
Nucléoles normaux.	Détermine si les nucléoles sont petits et à peine visibles ou plus gros, plus visible et plus abondante.	0-10
Mitoses	Décrit le niveau d'activité mitotique (reproduction cellulaire).	0-10
Classe	2 classes (2 pour bénignes, 4 pour malignes)	2/4

Tableau. **II.12**. Description des attributs de l'ensemble de données sur les maladies cardiaques.

Caractéristique	Description	Interval
Âge	Âge en années	0/1
Sexe	Sexe du sujet [femme=0, homme=1]	0/1
Cp	Type de douleur thoracique	0/1
Trestbps	Glycémie au repos	0/1
Chol	Cholestérol sérique	0/1
Fbs	Glycémie à jeun [vrai, faux]	0/1
Restecg	Résultat ECG au repos[normal, anomalie]	0/1
Thalach	Fréquence cardiaque maximale atteinte	0/1
Exang	Angine induite par l'exercice [non, oui]	0/1
Oldpeak	Dépression du segment ST induite par l'exercice par rapport au repos	0/1
Slope	Pente ou pic d'exercice Segment ST	0/1
Ca	Nombre de vaisseaux principaux colorés par fluoroscopie	0/1
Thal	Type de défaut	0/1

Smok	fumer (1 = oui ; 0 = non)	0/1
Thalrest	fréquence cardiaque au repos	0/1
tpeakbpd	pression artérielle maximale à l'exercice	0/1
trestbpd	tension artérielle au repos	0/1
xhypo	(1 = oui ; 0 = non)	0/1
rldv5e repos	hauteur à l'exercice de pointe	0/1
restef	fraction d'éjection du raidonuclide (sp?) au repos	0/1
restwm	anomalie du mouvement de la paroi du repos	0/1
Num	Diagnostic de maladie cardiaque (Valeur 0 : < 50 % de rétrécissement du diamètre -- Valeur 1 : > 50 % de rétrécissement du diamètre)	0/1
Classe	2 Classes (0 : non malade, 1 : malade)	0/1

Tableau . II.13 Description des attributs de l'ensemble de données sur les maladies des patients hépatiques

Attribut	Description	Interval
Âge	Âge des participants.	[04–90]
Genre	Sexe « féminin ou masculin » du patient	0–1
TB	Bilirubine totale	[0.4-75]
DB	Bilirubine directe	[0.1-19.7]
AlkPhos	Phosphatase alcaline.	[63-2110]
Sgpt	Alanine Aminotransférase	[10-2000]
Sgot	Aspartate aminotransférase.	[10-4929]
TP	Protéines Totales	[2.7 - 9.6]
ALB	Albumine	[0.9-5.5]
Ratio A / G	Ratio albumine et globuline	[0.3 - 2.8]
Classe	2 classes (teste positive: 1, teste negative: 0)	0–1

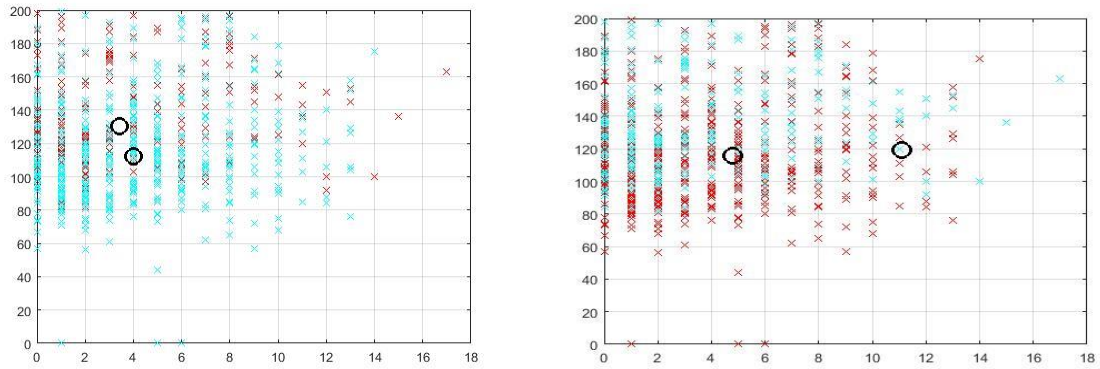
Tableau. II.14. Description des attributs de l'ensemble de données PID sur les maladies des patients diabétiques

Attribut	Description	Interval
Preg	Nombre de fois enceinte.	0–17
Glucose	Concentration de glucose plasmatique 2 h dans un test de tolérance au glucose par voie orale.	0–199
BP	Pression artérielle diastolique (mm Hg).	0–122
ÉpaisseurPeau	Épaisseur du pli cutané du triceps (mm).	0–99
Insulin	Insuline sérique de 2 heures (µlU/mL).	0–846
BMI	Indice de masse corporelle (kg/m2).	0–67.1
DPF	Fonction pedigree du diabète.	0.078–2.42
Age	Âge des participants.	21–81
Classe	2 classes (teste positive: 1, teste negative: 0)	0-1

II.7.2.2 Stratégie de simulation

Dans l'étape d'initialisation, nous avons donné au processus les bases de données et le nombre correspondant de classes. Ensuite, les quatre algorithmes initialisent aléatoirement les centres de classe dans l'espace de recherche. Par la suite, ils calculent la distance euclidienne (la fonction objective) entre les centres de classe et les instances.

Après cela, selon la meilleure solution obtenue, la classification est effectuée. La figure II.15 explique le processus de clustering.



°: centre de classe; * bleu: instances de classe 1; * rouge: instances de classe 2

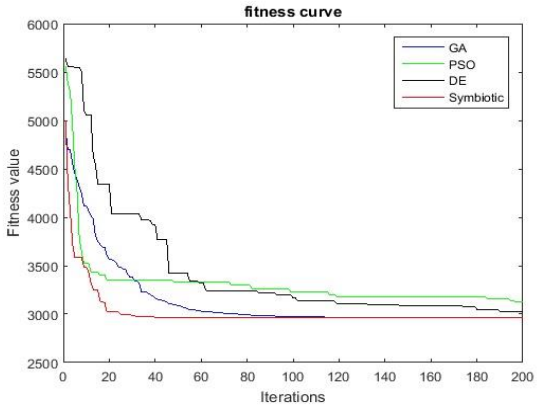
Figure II.15. Processus de regroupement de l'ensemble de données sur le diabète (PID)

II.7.2.3 Résultats de la classification

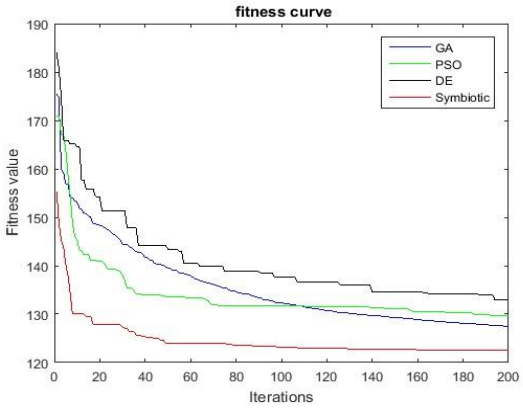
Les résultats ont été obtenus après 200 itérations des quatre algorithmes. Pour afficher des résultats plus détaillés, nous utilisons les courbes de fitness et les histogrammes de classification. Pour plus d'estimation des résultats, la validation est effectuée pour les quatre méthodes par la matrice de confusion et les courbes ROC.

II.7.2.3.1 Courbes de la fonction de fitness

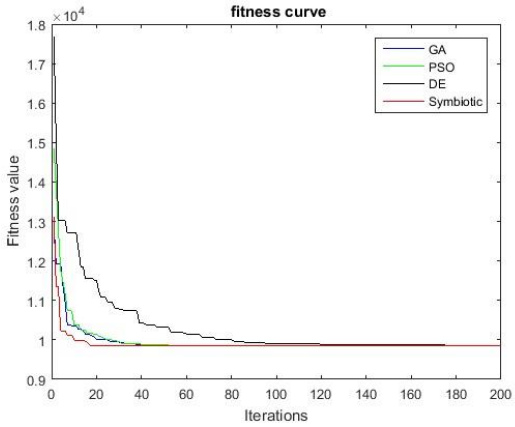
Dans cette section, nous avons évalué la fonction de fitness pour les quatre algorithmes sur les quatre bases de données précédemment décrites. Le but de cette évaluation est de montrer la convergence de SOS comparativement à d'autres méthodes.



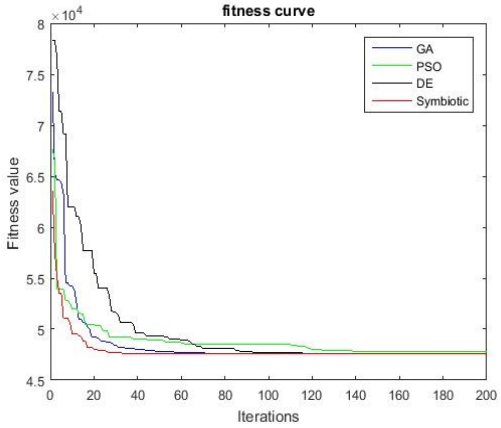
a : BC database



b : HD database



c : LD database



d : PID database

Figure II.16 : Courbes de la fonction Fitness pour les 4 algorithmes de classification

a. Base de données BC, b. Base de données HD, c. Base de données LD, d. Base de données PID

Les figures II.16a, II.16b, II.16c et II.16d représentent les variations de la fonction fitness. L'algorithme SOS est plus efficace et effectue une convergence plus rapide que les autres méthodes. Cela prouve la bonne qualité de la solution SOS par rapport aux autres métaheuristiques utilisées.

II.7.2.3.2 Histogrammes de classification

Dans cette étude, nous avons appliqué les quatre algorithmes pour la tâche de classification en utilisant les quatre bases de données pour découvrir si chaque algorithme classe chaque instance dans la classe appropriée. Pour afficher le résumé avec l'analyse des données, nous utilisons le tableau II.15 suivant :

Tableau . II.15. Distribution du nombre d'instances de chaque classe par les quatre algorithmes

		Le résultat de classification			
		GA	DE	PSO	**SOS**
BC base de données	**Class1 n=444 Bénin**	n=450	n=450.	n=450	**n=450**
	Class2 n=239 malin	n=233	n=233	n=233	**n=233**
HD base de données	**Class1 n=23 anormal**	**n=24**	n=53	n=25	n=27
	Class2 n=57 normal	**n=56**	n=27	n=55	n=53
LD base de données	**Class1 n=145 pas de désordre**	n=283	n=64	n=64	**n=71**
	Class2 n=200 désordre	n=62	n=281	n=281	**n=274**
PID data base	**Class1 n=500 pas infecté**	n=496	n=273	n=496	**n=500**
	Class2 n=268 infecté	n=272	n=495	n=272	**n=268**

Dans les figures II.17 et II.18 suivantes, nous représentons les données classées sous forme d'histogrammes pour donner une visibilité significative afin de faire une bonne comparaison entre les méthodes.

Nous n'avons sélectionné que deux bases de données qui sont le cancer du sein et les maladies cardiaques qui contiennent plusieurs fonctionnalités.

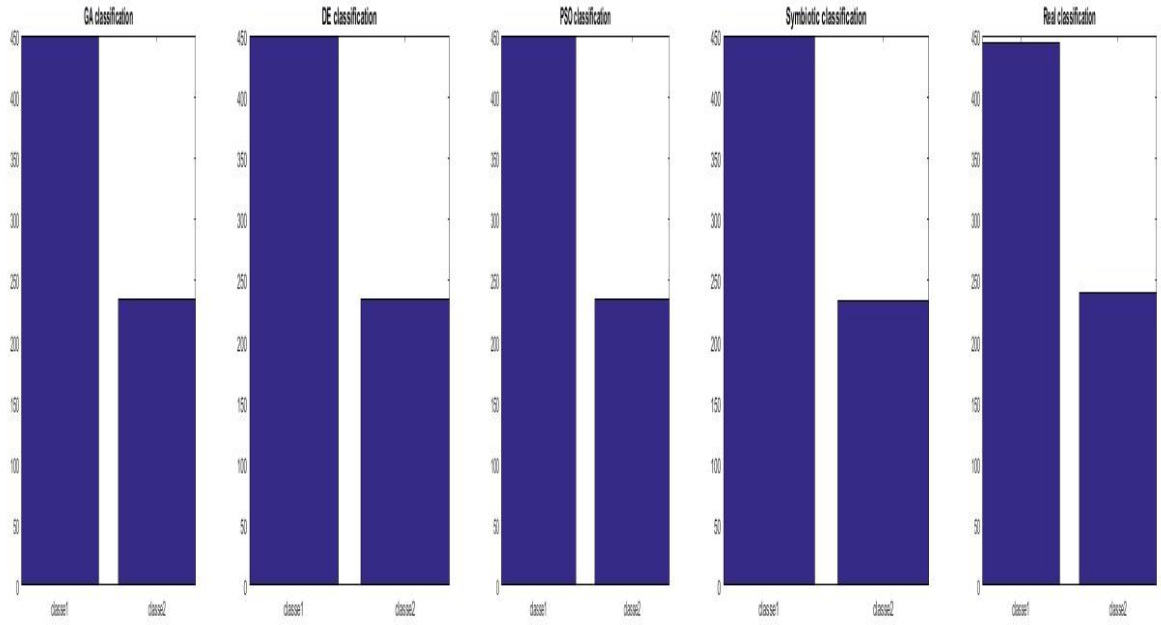


Figure. II.17. Histogrammes des classifications dans la base de données du BC

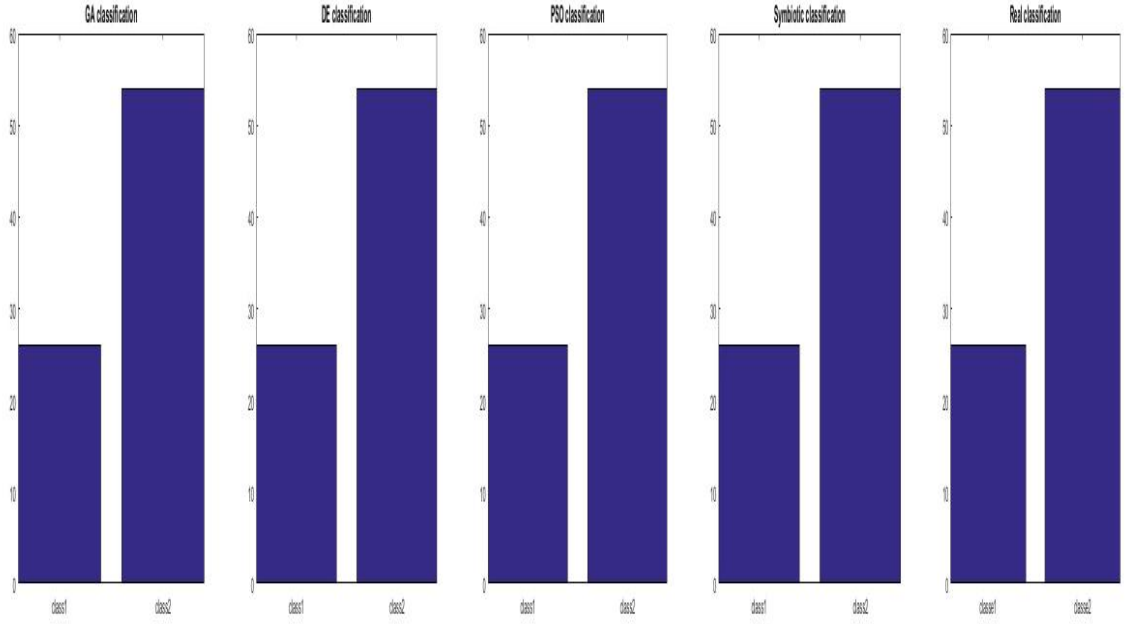


Figure II.18. Histogrammes des classifications dans la base de données HD

D'après les résultats présentés sur les différents histogrammes de classification ci-dessus, nous signalons :

- Il est clair que (voir le tableau II.15) la classification de l'histogramme SOS est presque identique au résultat réel.
- SOS a donné une très bonne classification car seulement quelques éléments ont été mal classés, alors que les autres méthodes génèrent davantage de nombres mal classés.

II.7.2.3.3 Validation de la classification à l'aide de la matrice de confusion

Pour une appréciation plus qualitative des résultats de classification par SOS et d'autres algorithmes utilisés dans ce travail de recherche, nous avons présenté des matrices de confusion (MC). La diagonale de chaque MC donne la distribution des instances dans la classe correspondante [106]. Le taux de précision (TP) ou précision : déduit du MC, montre l'ensemble des observations de bonne classification, et il est défini par l'équation suivante [106]:

$$TP=\sum_{i=1}^{n} diag(CM_i)/\sum_{i=1}^{n} CM_i \quad \text{(Eq.19)}$$

Où :

$\sum diag\ (MC_i)$: représente la somme des diagonales de la matrice de confusion des instances bien classées.

$\sum MC_i$: c'est la somme de tous les coefficients de la matrice de confusion.

n : nombre d'instances de la base de données.

Nous avons appliqué la matrice de confusion pour comparer les performances de SOS avec les autres méthodes citées, ainsi qu'avec GWOTS qui est une méthode hybride combinant l'optimiseur de loup gris (GWO) et la recherche Tabou (TS) [107]. Le taux de classification précis pour chaque base de données est enregistré dans le tableau II.16.

Tableau. II.16. La matrice de confusion Résultats pour les quatre bases de données utilisant les quatre métaheuristiques

	Le taux de précision (en%) des méthodes				
	GA	**DE**	**PSO**	**GWOTS**	**SOS**
BC base de données	**96.48**	95.31	95.75	95.422	**96.48**
HD base de données	65.25	63.25	64.25	**66.79**	**63.75**
LD base de données	52.75	54.65	49.96	57.97	**61.57**
PID base de données	39.32	39.18	39.19	**79.63**	**60,10**

- ***Base de données sur le cancer du sein***

D'après les résultats rapportés dans le tableau II.16 concernant cette base de données, nous notons que les cinq algorithmes ont donné des résultats acceptables mais SOS et GA ont donné des résultats identiques et meilleurs qui sont de 96,48%.

- ***Base de données sur les maladies cardiaques***

Nous notons à partir du tableau II.16 que l'algorithme SOS a donné une précision de 63,75%. Ce n'était pas la meilleure par rapport à ce que le GA et le PSO ont trouvé, mais SOS reste la meilleure méthode, contrairement aux autres algorithmes où les résultats varient entre 28% et 63 ou 65,25%. Nous concluons que SOS est crédible par rapport aux autres qui n'ont pas un comportement stable. La méthode GWOTS avait un taux de précision légèrement supérieur qui est de 66,79% par rapport à notre méthode. Mais il faut savoir que c'est une méthode hybride et elle combine l'efficacité des deux métaheuristiques (GWO et TS).

- ***Base de données sur les troubles hépatiques***

Pour cette base de données difficile, SOS est nettement meilleure que les autres métaheuristiques ; trois d'entre eux (GA, DE et GWOTS) ont des résultats moyens, mais PSO est le pire car elle nous a donné moins de 50%.

- ***Base de données sur le diabète***

En utilisant cette base de données très complexe, seul SOS donne un bon résultat par rapport aux trois autres méthodes (GA, DE et PSO), elles ont obtenu un taux de classification inférieur à 50% comme résultat de classification, mais GWOTS donne le meilleur résultat pour la même raison.

II.7.2.3.4 Validation de la classification à l'aide des courbes ROC

Pour visualiser clairement les performances des modèles utilisés, nous avons tracé les courbes ROC, qui utilisent la matrice de confusion pour la comparaison des modèles avec la classification réelle en faisant varier le seuil de 1 à 0 pour chaque cas [108].Si les courbes de résultats sont supérieures à la diagonale (50%), alors l'algorithme utilisé donne une bonne classification, sinon cela signifie sous la diagonal et la méthode utilisée donne un résultat incohérent.Les résultats des courbes ROC montrent la classification de chaque instance dans la classe réelle correspondante en utilisant les mêmes bases de données qui sont dans les figures suivantes.

- ***Base de données sur le cancer du sein (BC)***

L'algorithme SOS et GA ont donné une excellente classification mais les autres algorithmes DE et PSO étaient moins significatifs, comme le montre la figure II.19.

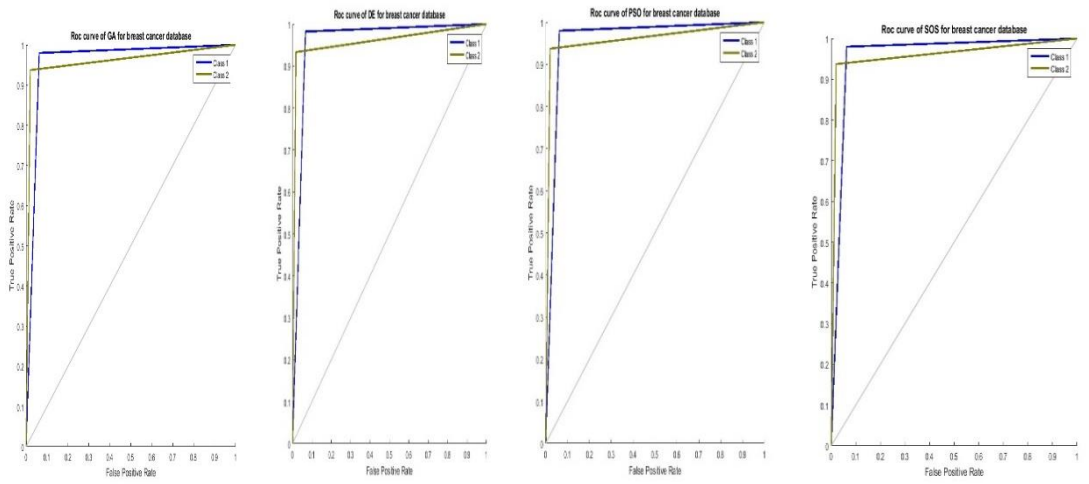


Figure II.19. Courbes ROC de la base de données BC pour les 4 algorithmes

- ***Base de données sur les maladies cardiaques (HD)***

Bien que SOS soit la deuxième méthode de classification de cette base, sa courbe ROC garde la même stabilité lors de plusieurs exécutions dans les mêmes conditions. Au contraire, les trois autres méthodes sont instables comme le montrent les figures II.20 et II.21 qui indiquent que la courbe de classification est sous la diagonale.

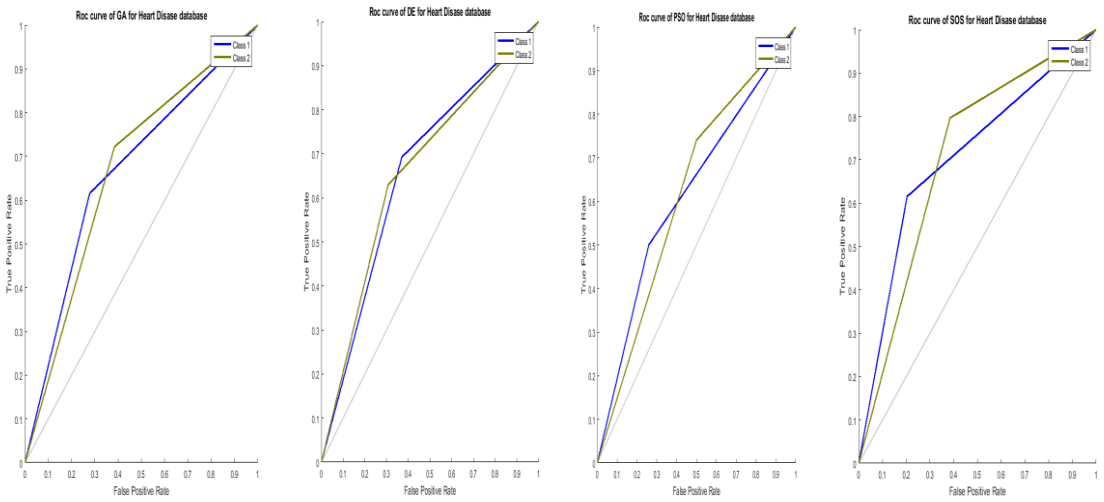


Figure II.20. Courbes ROC de la base de données HD pour les 4 algorithmes

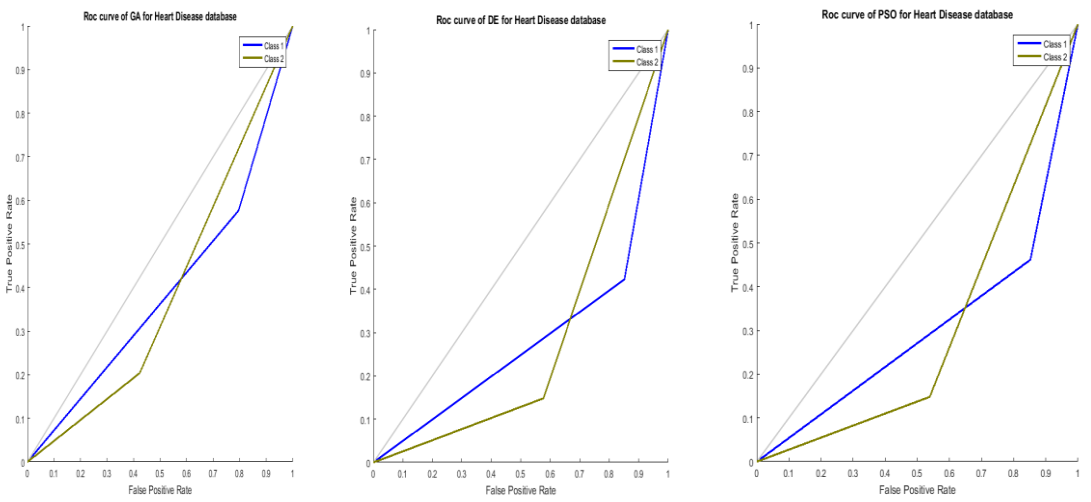


Figure II.21. Les courbes ROC de GA, DE et PSO exécutées sur la base de données HD donnent des résultats incohérents

- ***Base de données sur les troubles hépatiques (LD)***

Pour cette base de données complexe, SOS nous a donné de bons résultats par rapport aux autres méthodes clarifiées dans la figure II.22.

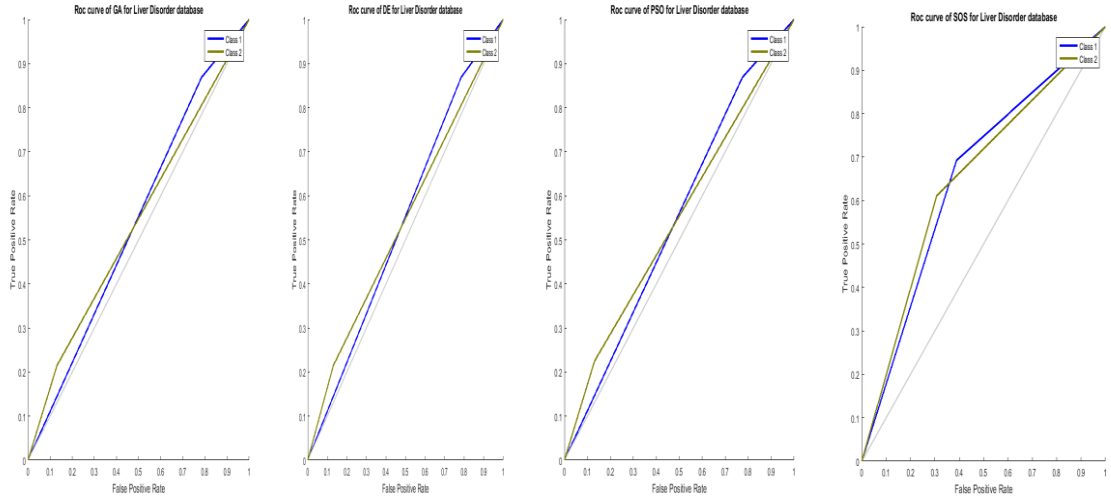


Figure II.22 Courbes ROC de la base de données LD pour les 4 algorithmes

- ***Base de données sur le diabète (PID)***

En utilisant cette base de données très complexe et difficile, nous avons constaté que SOS donnait de bons résultats correspondant à la complexité de cette base de données, mais les autres

méthodes nous donnaient une mauvaise classification sous la diagonale comme le montre la figure II.23.

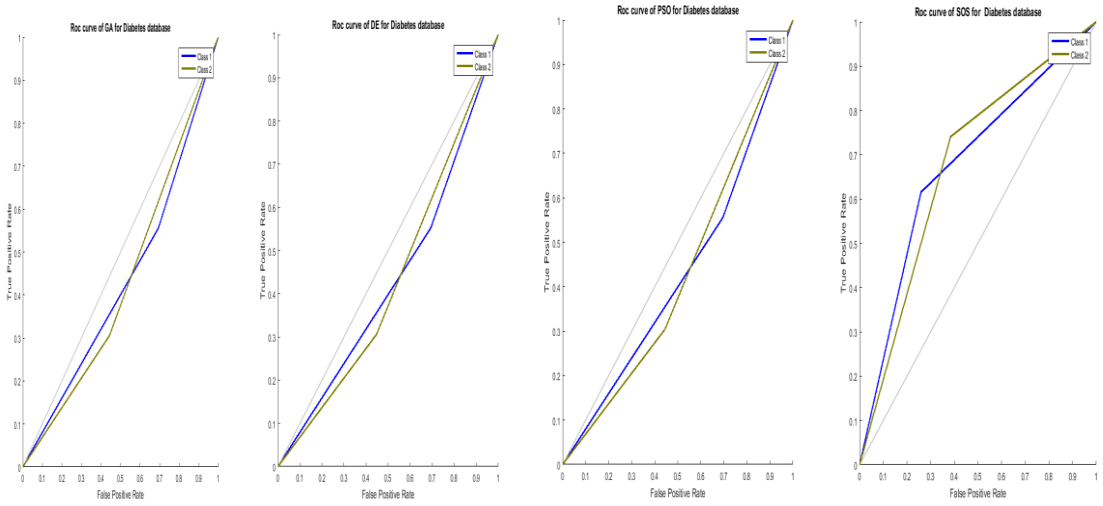


Figure II.23. Courbes ROC de la base de données du diabète PID pour les 4 algorithmes

II.8. Conclusion

Dans ce chapitre, nous avons proposé une nouvelle méthode de clustering supervisé concernant l'exploration de données. Notre technique proposée est basée sur l'algorithme SOS. Après la première évaluation expérimentale du SOS avec plusieurs fonctions de benchmark, SOS est régulièrement classé premier en termes de qualité de résultat par rapport aux autres algorithmes tels que PSO, ABS, CA, GWO, CDW-PSO et BOA, en particulier pour les problèmes à grande échelle.

Par conséquent, il a été très utile de l'appliquer sur des fonctions multimodales à haute dimension pour prouver que l'algorithme de recherche d'organismes symbiotiques (SOS) est une approche très intéressante pour l'exploration de données.

Pour enrichir nos choix théoriques, nous avons appliqué l'algorithme SOS au domaine médical. A travers les expériences utilisées, il était clair que la méthode symbiotique présente par rapport aux autres méthodes un grand aspect d'exploitation lui permettant d'avoir la solution optimale dans la plupart des cas pratiques réalisés.

À partir des courbes de la fonction objective, nous avons constaté que l'algorithme SOS converge rapidement vers la solution optimale et reste stable dans cette valeur contrairement aux autres algorithmes. Bien que SOS fasse ses premiers pas et que seule sa version de base ait été utilisée, contrairement aux autres algorithmes GA, PSO et DE dont leurs versions améliorées ont été utilisées, les résultats obtenus montrent que SOS est le meilleur algorithme de convergence.

Statistiquement, selon la matrice de confusionles taux de précision de classification obtenus par SOS étaient plus élevés que ceux des autres algorithmes dans les quatre ensembles de données.

Bien que les résultats obtenus aient prouvé l'efficacité de notre méthode, ils présentent encore quelques limites. Lorsque les attributs des bases de données sont très similaires, la méthode SOS converge vers l'optimum local.

A cet effet, dans le chapitre suivant nous allons definir formellement et valider une nouvelle version de SOS appelée VSOS (Velocity Symbiotic Organisms Search).

Chapitre III
Contribution 2 : Modélisation et validation d'une version améliorée de SOS

III.1 Introduction

Dans le domaine médical la prise de décision est onéreuse cela exige la meilleure méthode possible pour la classification et l'aide à la décision prédictive. La méthode basée sur la recherche d'organismes symbiotiques (SOS : symbioticorganismssearch) est l'une des métaheuristiques les plus importantes et les plus puissantes en raison de sa simplicité et de sa robustesse [10] [11]. Cependant, elle souffre de la convergence prématurée, cela se présente par le phénomène d'exploration qui se manifeste dans certain cas par une convergence vers l'optimum local au lieu de l'optimum global. Pour pallier à cette lacune, nous proposons une version améliorée de SOS appelée Vélocité Recherche d'Organismes Symbiotiques (VSOS : Velocity Symbiotic Organism Search) dans laquelle nous avons intégré la vitesse comme un nouveau paramètre d'optimisation.

Les réseaux neuronaux ont été aussi utilisés avec succès dans l'exploration de données pour diverses applications. Ils ont également été combinés avec des techniques métaheuristiques pour fournir des résultats supérieurs. Une contribution basée sur le couplage ou la combinaison de VSOS avec le réseau de neurone artificiel pour une classification rapide et optimale est réalisée.

Ce chapitre est organisé comme suit : la deuxième section contient une descriptiondes fondements théoriques de VSOS comme un modèle très efficace pour une analyse de petit ou un large volume de données. La troisième section de ce chapitre présente notre seconde contribution sous la forme de VSOS combiné avec MLP. Dans la quatrième section, nous développons une évaluation expérimentale de notre algorithme sur des fonctions tests et de classification de données en utilisant deux bases de données médicales et une troisième concernant plusieurs types de fleures.

III.2 Vélocité Recherche d'Organismes Symbiotiques (VSOS)

Pour résoudre la convergence prématurée de la méthode SOS, nous proposons l'ajout d'un nouveau terme, appelé vitesse dans les équations de mise à jour d'origine. Avec ce terme, les solutions peuvent aller plus loin que l'ancienne position, et assurer une exploration efficace de l'espace de recherche, position au déplacement d'un organisme nécessite une vitesse pour passer de sa position actuelle à une nouvelle. Le terme de vitesse est intégré dans la première et deuxième phase du SOS. Notre algorithme proposé est détaillé comme suit :

III.2.1 Phase de mutualisme

Les organismes vivants participent à une relation mutuelle dans le but d'augmenter l'avantage de la survie mutuelle dans l'écosystème.

Comme mentionné ci-dessus, dans la version originale de la méthode SOS, de nouveaux organismes candidats sont obtenus sur la base d'une symbiose mutuelle, en utilisant les équations. 13 et 14 dans le chapitre II.

Dans notre méthode VSOS proposée, un terme de vitesse (*V*) [100] est ajouté dans les équations 13 et 14, pour les solutions i et j. Les nouvelles équations obtenues sont 22 et 23.

$$V_i = w \times v_i \quad \text{(Eq.20)}$$

$$V_j = w \times v_j \quad \text{(Eq.21)}$$

$$X_{new1} = X_i + \lambda \times (X_{best} - Vecteur\_Mutual \times BF1) + V_i \quad \text{(Eq.22)}$$

$$X_{new2} = X_j + \beta \times (X_{best} - Vecteur\_Mutuat \times BF2) + V_j \quad \text{(Eq.23)}$$

$$v_i = X_{new1\_i} - X_i \quad \text{(Eq.24)}$$

$$v_j = X_{new2\_j} - X_j \quad \text{(Eq.25)}$$

V_i et V_j se réfèrent à la vitesse actuelle des organismes X_{new1_i} et X_{new2_j}.

Le processus commence avec des valeurs aléatoires de $v_i$et v_j. L'Eq. 20 et 21 sont la position de mise à jour de la vitesse et l'égaliseur. 24 et 25 sont les équations de mise à jour de la vitesse. Le coefficient d'inertie*w* est utilisé pour contrôler le terme de vitesse. Afin de faire passer l'algorithme de l'exploration à l'exploitation, nous utilisons des valeurs réduites de *w*, en utilisant l'équation 26.

$$w = (w_{min} - w_{max}) \times \frac{iter}{Max\_iter} + w_{max} \quad \text{(Eq.26)}$$

Où, w_{max} et w_{min} sont fixés expérimentalement à 0,7 et 0,1 respectivement.

La figure III.24 illustre un exemple de mise à jour de la position d'une solution candidate, en phase de mutualisme, à l'aide des méthodes SOS et VSOS. Il est clair que le terme ajouté aide la solution à aller plus loin dans l'espace de recherche et à explorer différentes régions de l'espace de recherche.

Avec le coefficient d'inertie, qui prend des valeurs décroissantes tout au long du processus d'optimisation, l'algorithme procure un bon équilibre entre l'exploration et l'exploitation.

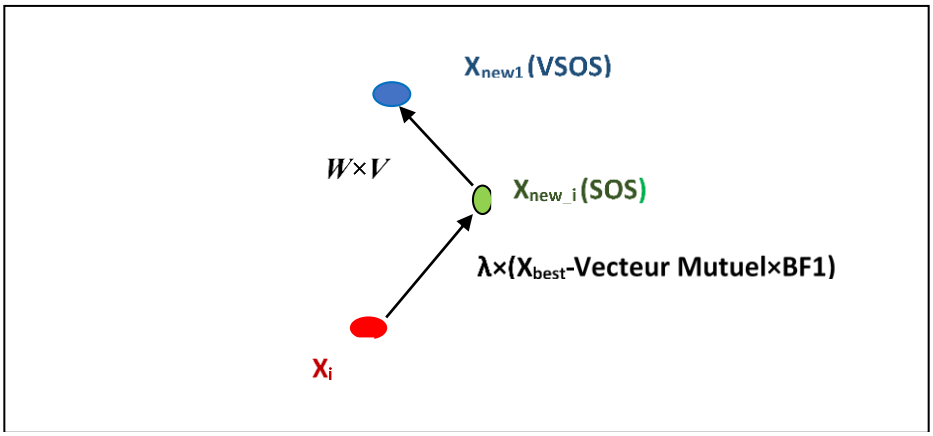


Figure III.24. La phase de mutualisme modifiée

III.2.2 Phase de commensalisme

La même étape dans cette phase, avec le terme de vitesse proposé, Eq. 16 est mis à jour comme suit :

$$V_i = w \times v_i \quad \text{(Eq.27)}$$

$$X_{new\_i} = X_i - \alpha \times (X_{best} - X_j) + V_i \quad \text{(Eq.28)}$$

$$v_i = X_{new\_i} - X_i \quad \text{(Eq.29)}$$

Le terme de vitesse ajouté *V* peut être vu comme un terme physique qui représente la tendance de l'organisme à suivre sa direction actuelle.

III.2.3 Phase de parasitisme

Dans la phase de parasitisme, le terme de vitesse n'est pas utilisé car la nouvelle solution est obtenue aléatoirement à partir de l'espace de recherche.

Cependant, pour améliorer davantage les performances SOS, nous avons utilisé la phase de parasitisme modifiée de l'algorithme SOS proposé dans [116], les auteurs ont proposé une phase de parasitisme modifiée, afin d'améliorer les performances de la méthode SOS.

La phase de parasitisme suggérée contient deux sous-phases, la phase de parasitisme d'origine et la phase de cleptoparasite. L'algorithme bascule de manière aléatoire entre eux.

Algorithme : la modification de la phase parasitisme

Si rand <0.6
    La phase originale de parasitisme
Si non
    La phasecleptoparasite
Fin Si

La phase cleptoparasitelimite le comportement des corbeaux (développé dans l'algorithme de recherche de corbeaux) et s'exprime comme suit :

$$X_{cleptoparasite_{new}} = X_{cleptoparasite} + \alpha(X_{best} - X_{cleptoparasite}) \quad \text{(Eq.30)}$$

Où $X_{cleptoparasite\_new}$ est considéré comme l'hôte d'un autre organisme, α est un facteur de pondération et est fixé à l'aide de l'équation suivante :

$$\alpha = rand\,(-1, 1) \times fl$$

Où fl représente la durée du vol, il est fixé à 2.

L'organigramme de la figure III.25 représente les principales étapes de notre algorithme VSOS proposé.

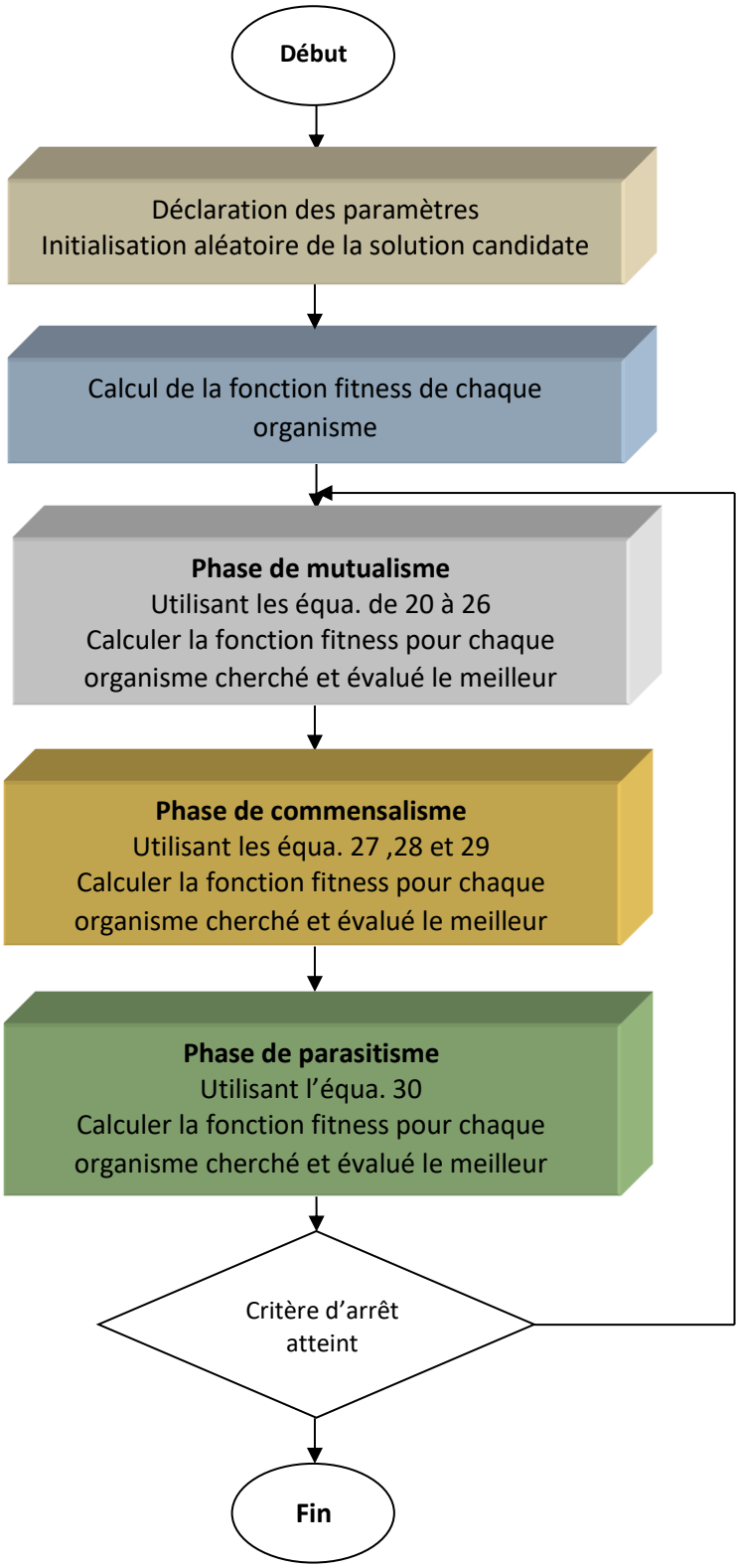


Figure. III.25. Organigramme de VSOS

III.2.4 Complexité informatique

La complexité de calcul du SOS original et de notre méthode VSOS est calculée en terme du nombre d'opérations mathématiques. Cela dépend aussi du nombre de solutions candidates (N), des dimensions (Dim) et du nombre maximum d'itérations (Max_iter). Pour la méthode SOS, la principale complexité de calcul réside dans les trois processus de mise à jour, elle peut être calculée à l'aide de l'Eq.32 :

$$O(SOS) = O(mutualism\ \ process) + O(commensalism\ \ process) + O(parasitism\ \ process) \quad \text{....(Eq. 31)}$$

Le nombre total d'opérations mathématiques nécessaires est de 14 opérations : 8 en phase de mutualisme, 3 en phase de commensalisme et 3 autres opérations en phase de parasitisme.

$$O(SOS) = 8 \times O(N.Dim.Max\_iter) + 3 \times O(N.Dim.Max\_iter) + 3 \times O(N.Dim.Max\_iter) \quad \text{(Eq.32)}$$

Pour notre méthode VSOS, avec le terme de vitesse ajouté, le nombre total d'opérations mathématiques est porté à 20 opérations mathématiques : 12 en phase de mutualisme, 5 en phase de commensalisme et 3 en phase de parasitisme.

$$O(VSOS) = 12 \times O(N.Dim.Max\_iter) + 5 \times O(N.Dim.Max\_iter) + 3 \times O(N.Dim.Max\_iter) \quad \text{(Eq. 33)}$$

La complexité de calcul de notre méthode VSOS est légèrement supérieure à celle de la méthode SOS, cependant, en termes de qualité de la solution, notre méthode VSOS est meilleure.

III.3. Perceptron multicouche (MLP) optimisé en utilisant l'algorithme VSOS

Dans cette section, notre algorithme VSOS proposé est utilisé pour la formation des MLP. VSOS est intégré au processus MLP (Multi-Layered Perceptron) pour trouver la combinaison optimale de pondérations et de biais qui minimisent l'erreur MLP.

III.3.1. Réseau de neurones perceptron multicouche

Le réseau de neurones artificiels (ou en Englais Artificial Neural Network : ANN) est un système composé d'au moins deux couches de neurones, et s'il contient des couches cachées, il est appelé MLP (Multi-Layered Perceptron). Un réseau MLP est constitué d'un ensemble fini de cellules (neurones) organisées en couches successives (voir figure III.26). La première couche avec les neurones R est appelée la couche d'entrée, la dernière est la couche de sortie avec les neurones S et les couches intermédiaires sont les couches cachées avec les neurones M. La notation **R-M-S** fera référence à la structure de MLP avec une seule couche cachée. Chaque cellule d'une couche est connectée à toutes les cellules de la couche suivante (réseau neuronal à alimentation directe).

Nous associons un coefficient de pondération (poids synaptique) à chaque connexion. Ces poids synaptiques sont notés W_1 et W_2 pour le réseau sur la figure III.26 [47] [109].

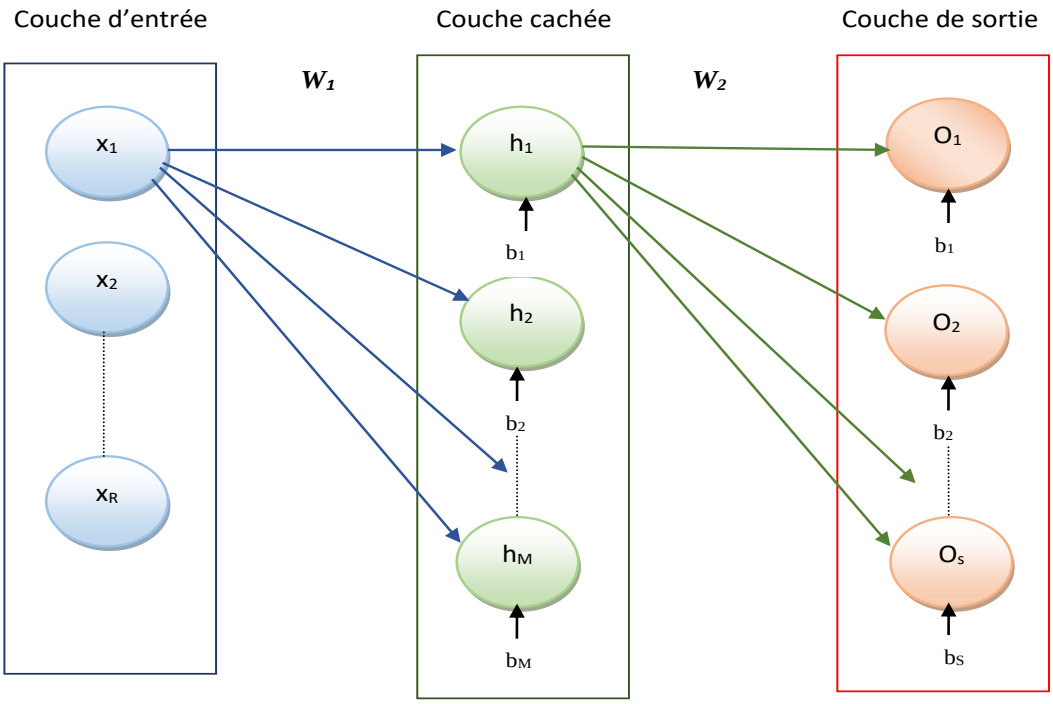


Figure III.26. Architecture du RNA-MLP (ANN-MLP)

La description formelle des composants ANN-MLP est la suivante :

III.3.1.1 Couche d'entrée

Différents signaux ; x_1, x_2,… x_R sont injectés dans les différents neurones de la couche d'entrée. Chaque lien est pondéré par un facteur ; w_1, w_2,..,w_R. Ce poids est adapté tout au long du processus d'apprentissage pour permettre au réseau de valoriser les résultats (en général il reste compris entre 0 et 1 ou -1 et 1) [47] [109].

III.3.1.2 Couche cachée

Dans ce niveau, la somme de tous ces signaux pondérés est calculée comme suit :

$$\sum_{i=1}^{R} w_i \times x_i \qquad \text{(Eq.34)}$$

Ensuite, un certain biais (b) est ajouté, il peut être vu comme un neurone externe supplémentaire, qui envoie systématiquement un signal 1 de poids b au neurone concerné. Une fois cette somme calculée, nous appliquons une fonction d'activation pour obtenir le signal de sortie de la couche cachée. Il existe plusieurs fonctions d'activation, les fonctions les plus utilisées sont le seuil, linéaire et sigmoïde [47] [109].

Dans cette approche, nous utilisons la fonction sigmoïde, et elle est définie comme suit :

$$f_{activation}(x) = \frac{1}{1 + exp^{-x}} \quad \text{(Eq.35)}$$

Par conséquent, la formule de sortie d'un neurone caché j, avec j = 1, 2, ... M, est donnée comme suit :

$$Oh_j = f_{activation}(\sum_{i=1}^{R}(w_{ij} \times x_i) + b_j) \quad \text{(Eq.36)}$$

II.3.1.3 Couche de sortie

Pour la couche de sortie (k = 1, 2, ... S), la formule de sortie des composants ANN-MLP est obtenue comme suit [47] [109]:

$$\tilde{y} = f_{activation}(\sum_{j=1}^{M}(w_{jk} \times Oh_k) + b_k) \quad \text{(Eq.37)}$$

S'il y a une erreur entre la sortie souhaitée (ỹ) et réelle (y), nous devons corriger les poids comme suit [47] [109]:

$w_i(l+1) = w_i\,(l) + \Delta w$ (Eq. 38)

$\Delta w = \alpha \times x_i \times er$ (Eq. 39)

Où l désigne le numéro d'itération, α est le taux d'apprentissage et er = ỹ-y représente l'erreur.

Le taux mesure la capacité du classificateur, la formule du taux de précision est définie par [109]:

Taux de classification (%) = (ñ/n) ×100 (Eq.40)

Où ñ est le nombre d'objets correctement classés et n indique le nombre d'entrainement.

III.3.2 Initialisation

Premièrement, tous les poids et biais sont initialisés au hasard.

$$X = \begin{bmatrix} w_{1,1}, w_{1,2}, & w_{1,3}, \dots w_{1,nw}, & \theta_{1,1}, & \theta_{1,2}, & \theta_{1,3}, \dots \theta_{1,nb} \\ w_{2,1}, w_{2,2}, & w_{2,3}, \dots w_{2,nw}, & \theta_{2,1}, & \theta_{2,2}, & \theta_{2,3}, \dots \theta_{2,nb} \\ & & \vdots & & \\ w_{N,1}, w_{N,2}, & w_{N,3}, \dots w_{N,nw}, & \theta_{N,1}, & \theta_{N,2}, & \theta_{N,3}, \dots \theta_{N,nb} \end{bmatrix}$$

Où *nw*, *nb* et *N* représentent respectivement le nombre de poids, de biais et la taille de la population. Le nombre de neurones d'entrée (*I*) est le nombre d'entités dans l'ensemble de données. Le nombre de neurones dans la couche cachée est calculé à l'aide de l'équation suivanteselon le théorème de Kolmogorov [110] : $H = 2 \times I + 1$.

Le nombre de neurones de sortie(O) est le nombre de classes dans l'ensemble de données. Ainsi, le nombre total de dimensions est :

$$Dim = (I \times H) + (H \times O) + H + O \quad \text{(Eq.41)}$$

Dans les couches cachées et en sortie, nous avons utilisé la fonction sigmoïde pour la validation. Nous avons choisi cette fonction, car c'est une fonction différentiable, nous pouvons donc trouver la pente de la courbe sigmoïde en deux points quelconques. De plus, ses valeurs sont dans l'intervalle [0,1], donc pour les modèles où il faut prédire une valeur de probabilité en sortie, si le bon choix.

III.3.3 Fonction de Fitness

Les solutions candidates sont ensuite évaluées avec une fonction de fitness, nous avons utilisé le critère de l'erreur quadratique moyenne (ou MSE : mean square error), pour mesurer la différence entre les valeurs réelles et souhaitées données par le neurone de sortie [109].

Le but de l'algorithme est de minimiser la valeur MSE, de sorte qu'une valeur MSE inférieure correspond à une meilleure valeur de fitness. La fonction de fitness est exprimée en (Eq.42) comme suit [109] :

$$ff = MSE = \frac{1}{n}\sum_{i=0}^{n}(\tilde{y} - y)^2 \quad \text{(Eq. 42)}$$

Oùnest le nombre d'échantillons dans l'ensemble de données d'apprentissage. La solution candidate qui donne la valeur de fitness minimale est considérée comme la meilleure solution globale trouvée.

III.3.4. Critère d'évaluation

Pour évaluer les performances du classificateur, plusieurs critères sont utilisés (voir le tableau III.17) [106]:

Tableau III.17 Critères d'évaluation des classificateurs

Critère	Formule
Exactitude	= (TP + TN)/(TN + TP + FN + FP)
Rappel	= (TP)/(FN + TP)
Spécificité	= (TN)/(TN + FP)
Précision	= (TP)/(FP + TP)
F-mesure	= (2 × Précision × Rappel)/(Précision + Rappel)

Où :

- Vrai positif (**TP** : True Positive) : l'observation est positive et devrait être positive.
- Faux négatif (**FN** : False Negative) : l'observation est positive, mais elle est prédite négative.
- Vrai négatif (**TN** : TrueNegative) : l'observation est négative et devrait être négative.

• Faux positif (**FP** : False Positive) : l'observation est négative, mais elle est prédite positive.

Le calcul de la matrice de confusion est illustré dans le tableau III.18 [106].

Tableau III.18 Matrice de confusion

		classeActuale		
		Négative	Positive	
classePredicté	Négative	**TN**	**FP**	Rappel
	Positive	**FN**	**TP**	Sensitivité
		Précision	Spécificité	Accuracy

Où

• Négatif : représente la 1ère classe (Classe 1)

• Positif : représente la 2e classe (Classe 2)

• Positif (P) : l'observation est positive.

• Négatif (N) : l'observation n'est pas positive.

L'organigramme et l'architecture de notre MLP-ANN basé sur l'approche VSOS est représenté dans les figures III.27 et III.28

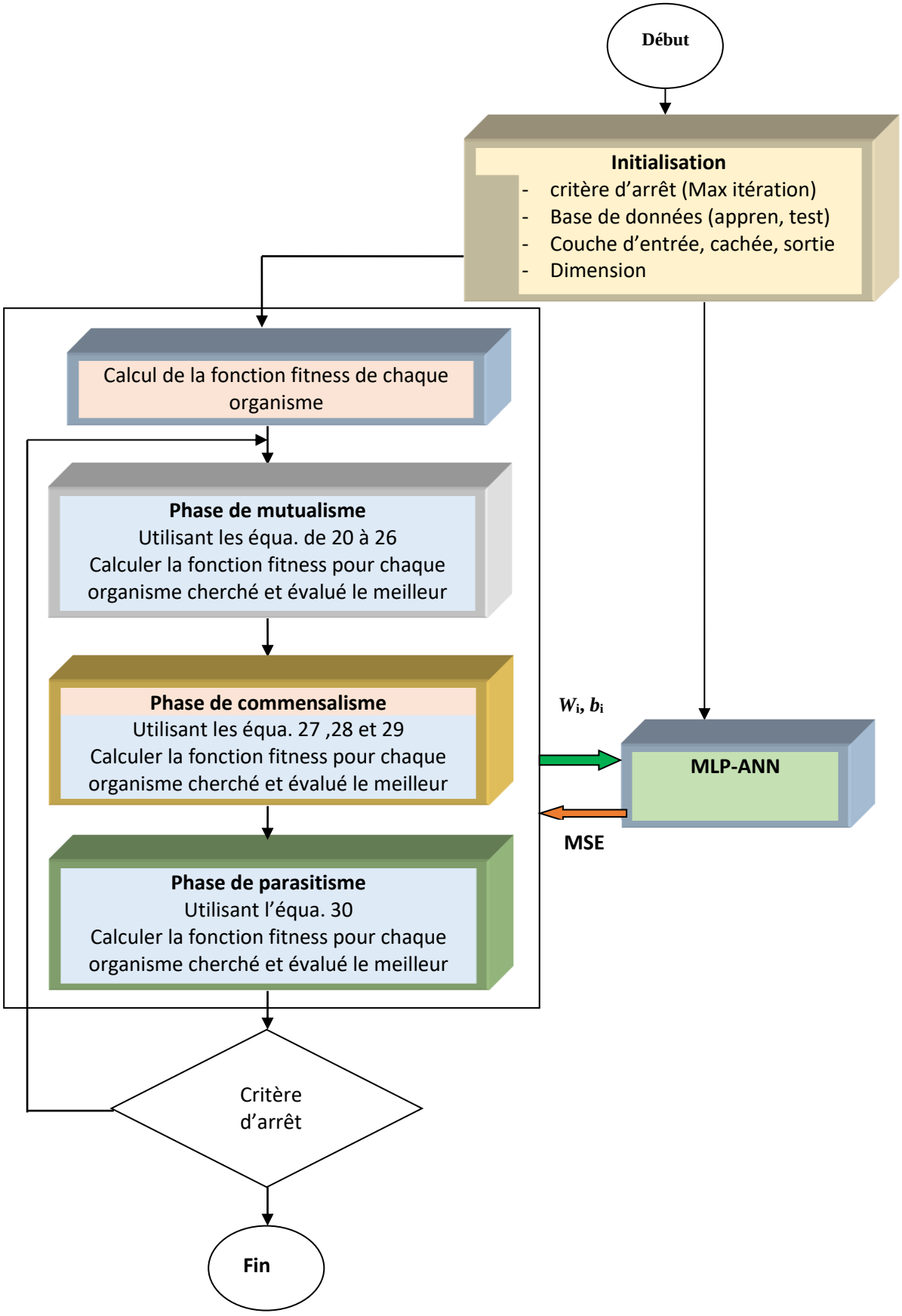


Figure III.27**.** Organigramme MLP-VSOS

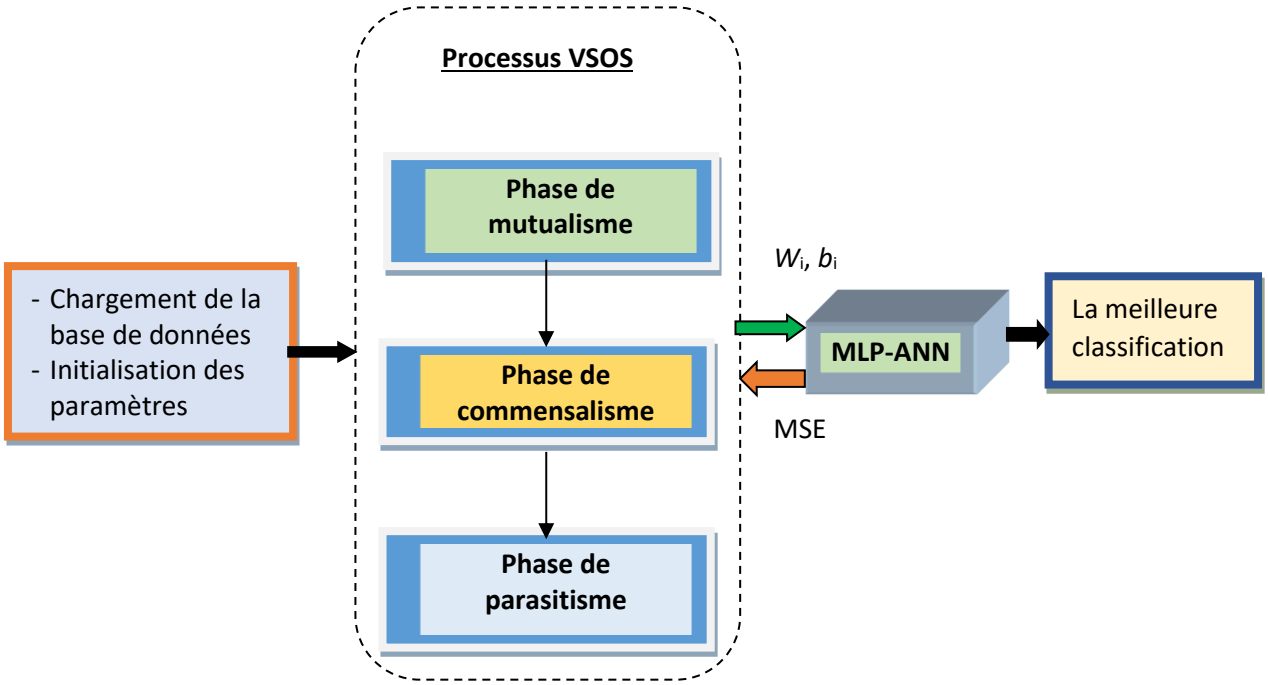


Figure III.28. Architecture de MLP-VSOS

III.4 Expérimentations et discussion

Dans cette section, les performances de l'algorithme VSOS proposé sont évaluées en deux phases. Tout d'abord, il est testé à l'aide de vingt-deux fonctions de référence bien connues [97]. Ensuite, la méthode MLP basée sur l'algorithme VSOS est évaluée par trois bases de données réelles et publiques.

III.4.1 Première expérience : Evaluation de VSOS

Les trois phases de VSOS sont exécutées pendant de nombreuses itérationspour explorer efficacement l'espace de recherche et converger progressivement vers la solution optimale globale. Le nombre maximal d'itérations Max_iter est fixé à 25, il est défini comme critère d'arrêt. En s'appuyant sur vingt-deux fonctions de référence classiquesdéfinies dans le CEC (Congress on Evolutionary Computation) 2005 du Congrès IEEE [111] sur le calcul évolutifsont utilisées dans la première expérience. Ils peuvent être regroupés en trois catégories ; les fonctions unimodales de grande dimension (F1−F07), les fonctions multimodales de grande dimension (F8 - F12) et les fonctions multimodales de faible dimension (F13 − F22). Les fonctions unimodales sont conçues pour tester la capacité d'exploitation des algorithmes d'optimisation. Les fonctions multimodales sont les fonctions les plus difficiles à optimiser car elles ont plusieurs optima locaux, et elles sont utilisées pour tester la capacité exploratoire des algorithmes. Une description détaillée de chaque fonction de test, y compris sa formulation mathématique, ses dimensions, la portée de la recherche et la solution optimale, est présentée à l'annexe A.

- ***Résultats d'évaluation de VSOS***

Les résultats de notre méthode VSOS sont comparés à ceux obtenus avec l'Optimisation pour essaim de particules (PSO) [100], l'algorithme d'optimisation d'Archimède (AOA) [112], Recherche de méduses (JS :Jellyfish Search) [113], les automates d'apprentissage basés sur l'optimiseur de loup gris (GWOLA : Grey wolf Optimizer Learning Automata) [114], l'algorithme d'optimisation de papillon (BOA : butterfly optimization algorithm) [102], le SOS original [10], SOS adaptatif (ASOS) [115], SOS amélioré (ESOS :Enhance SOS) [116]. Chaque algorithme est exécuté pour 25 analyses indépendantes, la valeur moyenne (Moy), l'écart type (EcTp) et le rang de chaque algorithme sont indiqués dans les tableaux III.19, III.20 et III.21.

Tableau. III.19 Résultats d'optimisation obtenus avec VSOS et d'autres métaheuristiques pour des fonctions unimodales de grande dimension

Fonctions		**PSO**	**GWOLA**	**BOA**	**AOA**	**JS**	**SOS**	**ESOS**	**ASOS**	**VSOS**
F1	Moy	0.3970	0	0,14262	1,46e-106	3,49e-19	0	4,51E-268	5,55e-150	**0**
	EcTp	0.3209	0	0,00335	7,32e-106	2,59e-19	0	0	1,95e-149	**0**
	Rang	7	1	6	4	5	1	2	3	**1**
F2	Moy	0.0352	0	0,03905	1,46e-115	6,03e-22	0	5,33E-261	6,33e-143	**0**
	EcTp	0.1624	0	0,01022	7,09e-115	5,70e-22	0	0	1,48e-142	**0**
	Rang	6	1	7	4	5	1	2	3	**1**
F3	Moy	1.6452e+3	0	0,16949	1,53e-81	0,08509327	0	4,37E-268	6,55e-150	**0**
	EcTp	640.9846	0	0,01199	7,66e-81	0,09387926	0	0	2,14e-149	**0**
	Rang	7	1	6	4	5	1	2	3	**1**
F4	Moy	2.1999	0	0,05345	1,19e-114	7,60e-16	0	5,07E-274	5,09e-145	**0**
	EcTp	1.7124	0	0,00214	5,90e-114	2,42e-16	0	0	1,90e-144	**0**
	Rang	7	1	6	4	5	1	2	3	**1**
F5	Moy	33.0197	25.1426	28,8594	27,1430594	13,9366588	28.8209	28,4366654	28,3338174	**0.0285**
	EcTp	31.9313	0.77188	0,02813	11,5592488	0,61417405	0.0334	0,29498086	0,32690671	**0.2e-03**
	Rang	9	3	8	4	2	7	6	5	**1**
F6	Moy	2.7682e+4	4.15835	6934,085	36,1946716	4091,24616	2.9951	4,14308165	6,37618827	**0.050**
	EcTp	4.9476e+3	0.66615	279,2301	34,9678973	868,632187	0.7141	0,97327362	1,02586945	**0.7e-03**
	Rang	9	4	8	6	7	2	3	5	**1**
F7	Moy	0.0251	1.94e-05	0,270496	0,00040563	0,00036777	1.6e-03	0,00351088	0,00321125	**0**
	EcTp	0.0170	1.37e-05	0,009268	0,00033179	0,00011317	2.2e-03	0,0025455	0,00205552	**0**
	Rang	8	2	9	4	3	5	7	6	**1**

Tableau. III.20 Résultats d'optimisation obtenus avec VSOS et d'autres métaheuristiques pour les fonctions multimodales de grande dimension

Fonctions		**PSO**	**GWOLA**	**BOA**	**AOA**	**JS**	**SOS**	**ESOS**	**ASOS**	**VSOS**
F8	Moy	-9.946e+3	-1,50e308	-3269,02	-4843,06868	-5480,05156	-6.58e+33	-6785,1914	-7430,20714	**-7.1555+44**
	EcTp	514.9218	0	384,6736	372,133733	115,473293	666.3787	710,20371	537,406386	**0.6061**
	Rang	4	**1**	9	8	7	3	6	5	**2**
F9	Moy	275.2287	**0**	275,7700	18,6050387	157,821279	**0**	0	86,5386576	0.0457
	EcTp	17.2402	**0**	17,69484	46,3014364	15,5642862	**0**	0	61,9827252	0.0514
	Rang	6	**1**	7	3	5	**1**	1	4	**2**
F10	Moy	13.4428	8.88e-16	20,06828	1,98E-05	19,3972766	8.8818e-16	4,58E-15	7,42E-08	**8.8818e-16**
	EcTp	0.7387	0	0,225689	2,67E-05	0,44255796	0	9,09E-15	7,02E-08	**0**
	Rang	5	**1**	7	4	6	**1**	2	3	**1**
F11	Moy	43.5010	0	29,95350	0,00994193	1,16815179	0	0	0,04391136	**0**
	EcTp	11.9431	0	0,559410	0,0411774	0,06518776	0	0	0,12255855	**0**
	Rang	6	**1**	5	2	4	**1**	**1**	3	**1**
F12	Moy	0.0456	0.01457	0,197116	0,01166533	8,45e-20	0.1628	0,2249084	0,18348552	**0.3e-03**
	EcTp	0.0902	0.00845	0,058056	0,0320942	8,18e-20	0.0616	0,10277782	0,07024651	**0.1e-03**
	Rang	5	4	7	3	**1**	6	9	8	**2**

Tableau. III.21 Résultats d'optimisation obtenus avec VSOS et d'autres métaheuristiques pour les fonctions multimodales de faible dimension

Fonctions		**PSO**	**GWOLA**	**BOA**	**AOA**	**JS**	**SOS**	**ESOS**	**ASOS**	**VSOS**
F13	Moy	0.9980	0.9980	0,998003	0,99800641	0,99800384	0.9980	0,99800384	0,99800384	**0.998004**
	EcTp	4.3578e-13	1.65e-10	5,59E-08	6,87E-06	2,01E-15	0	0	0	**0**
	Rang	3	4	5	6	2	1	1	1	**1**
F14	Moy	0.0036	5.475e-4	0,001664	0,00064368	0,00060837	3.4411e-04	0,000454	0,00030749	**3.075e-4**
	EcTp	0.0075	3.599e-4	0,001175	0,00012291	0,00036779	1.8314e-04	0,00034262	1,33E-19	**0**
	Rang	9	5	8	7	6	3	4	2	**1**
F15	Moy	-1.0316	-1.03162	-0,54502	-1,03162811	-1,03162845	-1.0316	-1,03162845	-1,03162845	**-1.000**
	EcTp	7.0682e-13	1.48e-06	0,329944	1,08E-06	6,41E-16	6.7987e-16	6,80E-16	6,80E-16	**0**
	Rang	6	8	9	7	2	3	4	4	**1**
F16	Moy	0.3979	0.39789	0,398713	0,39788844	0,39788736	0.4	0,39788736	0,39788736	**0.3979**
	EcTp	1.0120e-11	1.33e-05	0,001640	2,03E-06	0	0	0	0	**0**
	Rang	2	6	3	5	4	7	4	4	**1**
F17	Moy	3.0000	3.00002	3,049772	3,00000001	3	3.000	3	3	**3.0**
	EcTp	1.8965e-12	7.13e-06	0,138645	4,00E-08	8,60E-16	0.5368e-16	1,11E-15	7,90E-16	**0**
	Rang	6	8	9	7	4	2	5	3	**1**
F18	Moy	-3.8628	-3.8622	-3,63150	-3,86275915	-3,86278215	-3.8628	-3,86278215	-3,86278215	**-3.9**
	EcTp	1.0583e-7	1.153e-3	0,170873	5,10e-05	1,87e-15	2.2662e-15	2,27E-15	2,27E-15	**0**
	Rang	6	8	9	7	2	3	4	4	**1**
F19	Moy	-3.2602	-3.2504	-1,89728	-3,31404604	-3,32189167	-3.2839	-3,2696822	-3,24590357	**-3.320**
	EcTp	0.0606	0.05962	0,421318	0,02773574	0,0003203	0.0566	0,06023397	0,0582455	**0.1e-03**
	Rang	6	7	8	3	2	4	5	8	**1**
F20	Moy	-6.9258	-10.1517	-1,35731	-7,44497548	-5,07507009	-8.7258	-7,91007882	-9,83236248	**-10.150**
	EcTp	3.0358	0.000871	0,839147	2,25994264	0,09936182	2.3362	2,58276405	1,50197933	**0.000014**
	Rang	7	**2**	9	6	8	4	5	3	**1**
F21	Moy	-7.8961	-10.4015	-1,44006	-10,1858037	-5,08767183	-8.2768	-9,12727607	-9,98101115	**-10.4**
	EcTp	3.4918	9.84e-04	0,561386	0,69676571	2,84E-15	2.6576	2,31687193	1,46033708	**0**
	Rang	7	**2**	9	3	8	6	5	4	**1**
F22	Moy	-8.5370	-10.534	-1,72097	-10,2298128	-5,12848079	-9.0222	-9,67114117	-10,2683705	**-10.5**
	EcTp	3.3054	7.28e-04	0,595891	1,11788668	3,63E-15	2.4782	2,02346176	1,3401966	**0**
	Rang	7	**2**	9	4	8	6	5	3	**1**
Rang moyen		6.0909	3.0000	7.3182	4.5909	4.5000	3.0455	3.7727	3.8636	**1.8182**
Classement général		**8**	**2**	**9**	**7**	**6**	**3**	**4**	**5**	**1**

En ce qui concerne les résultats des tableaux III.19, III.20 et III.21, la méthode VSOS proposée surpasse la méthode SOS originale et toutes les autres métaheuristiques citées, et se classe première dans 19 fonctions : F1, F2, F3, F4, F5, F6, F7, F10, F11, F13, F14, F15, F16, F17, F19, F20, F21, F22. La méthode VSOS a également obtenu des résultats acceptables et elle est classée à la deuxième place juste dans 3 fonctions : F8, F9, F12. Ce qui prouve ses capacités d'exploitation et d'exploration. Dans la méthode VSOS proposée, un nouveau terme de vitesse est ajouté, pour améliorer la capacité d'exploration du SOS. Pour montrer la supériorité de notre méthode VSOS dans la phase d'exploration, les positions des solutions des candidats sont tracées, à différentes itérations : itération 1, 5 et 10, pour la fonction 22, qui est une fonction multimodale et qui a une solution optimale globale et deux optimums locaux. La figure III.29 montre les positions des optimums global et local.

Dans les figures III.30 et III.31, nous illustrons comment les positions des solutions candidates sont distribuées dans l'espace de recherche, avec les méthodes SOS et VSOS respectivement.

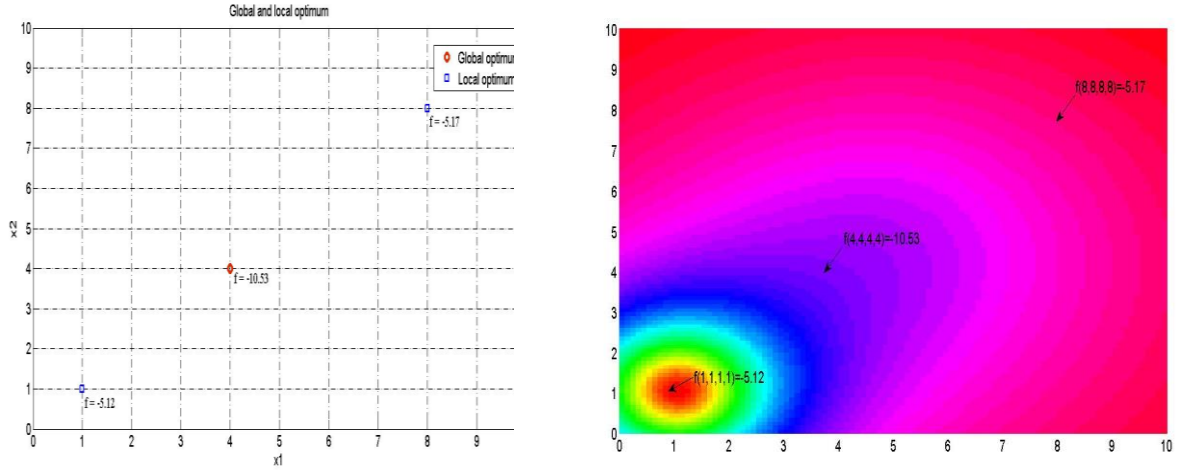


Figure III.29. La distribution de la solution optimale globale et locale pour F22

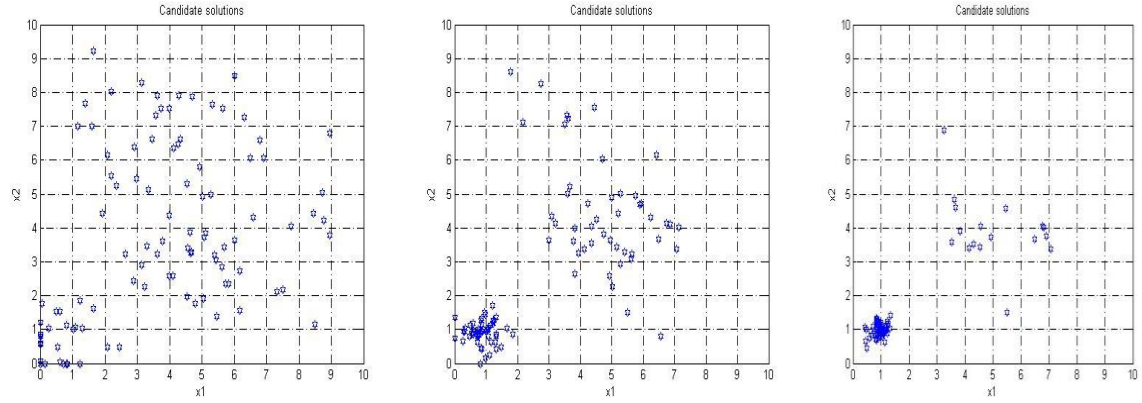


Figure. III.30 La répartition de la solution candidate dans l'espace de recherche en utilisant la méthode SOS pour les itérations 1, 5, 10

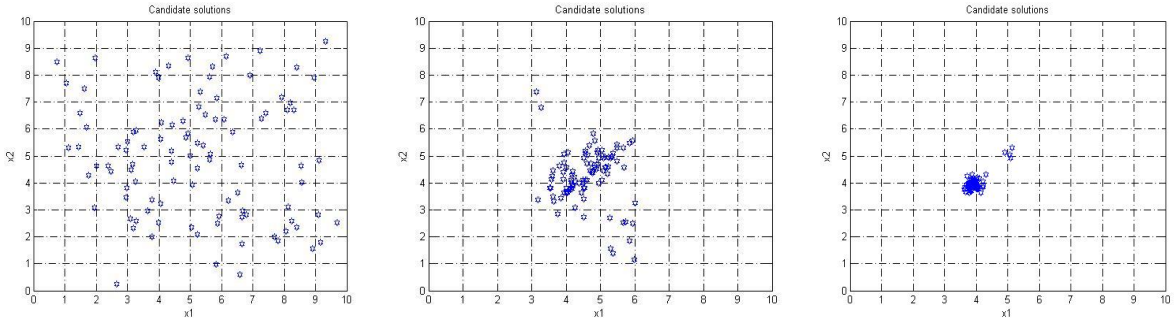


Figure. III.31 La distribution de la solution candidate dans l'espace de recherche en utilisant la méthode VSOS pour les itérations 1, 5, 10

Sur la figure III.30, la méthode SOS a été piégée dans un optimum local et n'a pas réussi à trouver la solution optimale globale. Sur la figure III.31, il est clair que la méthode VSOS a exploré l'espace de recherche de manière efficace et qu'elle a convergé vers la solution optimale globale. Ce qui prouve l'avantage du nouveau terme de vitesse, et confirme sa robustesse.

Pour obtenir une analyse graphique de notre méthode proposée, nous avons utilisé deux tracés : les courbes de convergence et les graphiques ANOVA (Analyse de Variation).

La figure III.32 illustre les courbes de convergence, obtenues avec notre méthode VSOS et les autres méthodes citées. Il est clair que l'algorithme VSOS a une vitesse de convergence plus rapide que les autres métaheuristiques, ce qui prouve sa robustesse et sa capacité de recherche.

La figure. III.33 présente le graphique ANOVA est un outil utile pour afficher visuellement la plage, la symétrie de distribution et la tendance centrale d'une distribution, elle nous permet également d'illustrer clairement la variabilité et la concentration de valeurs. Cinq nombres principaux sont représentés dans cet encadré ; la valeur minimale (la ligne dans une case du bas), la valeur médiane, la valeur supérieure (la ligne dans une case du haut), la valeur maximale et le marqueur (+) qui représente la plus grande valeur dans l'ensemble de données [32].

La distance entre la ligne supérieure de la boîte et la ligne inférieure représente l'écart type des 25 analyses indépendantes. D'après la figure III.33, il est évident que l'algorithme VSOS a un écart type plus faible, ce qui confirme sa stabilité.

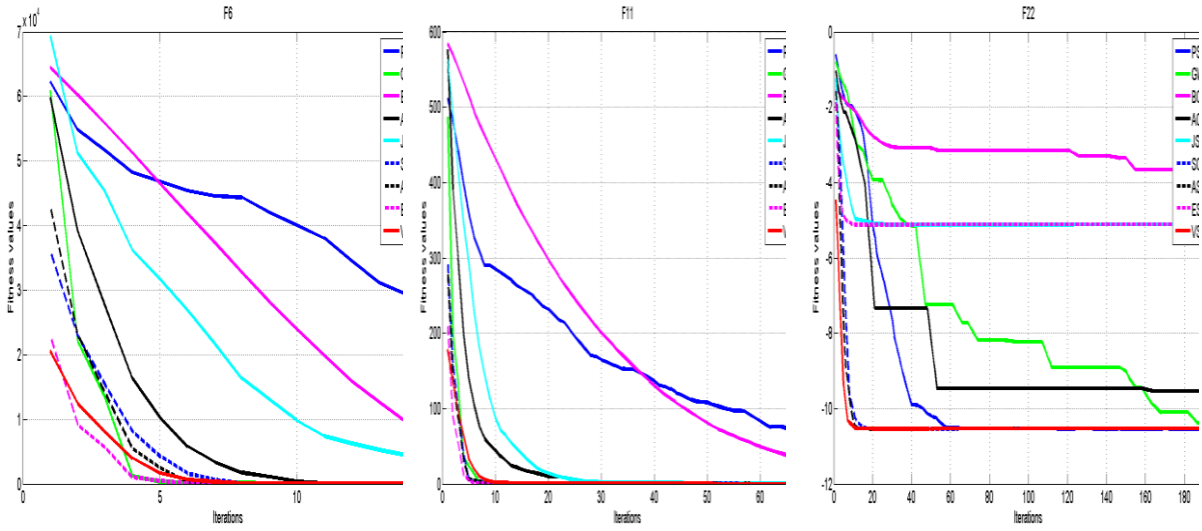


Figure. III.32 Courbes de convergence obtenue avec les différentes méthodes pour les fonctions F6, F11 et F22

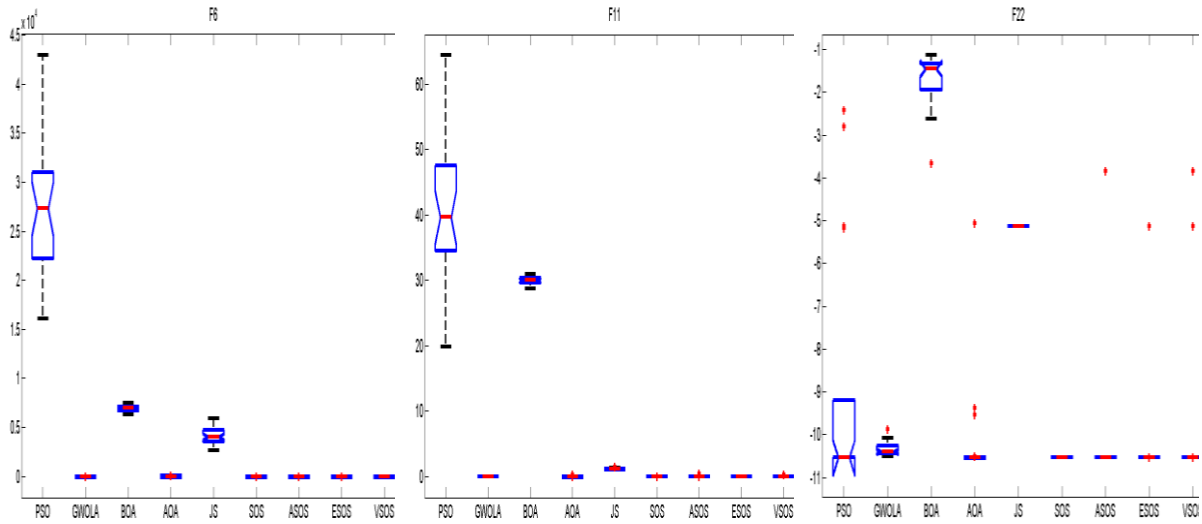


Figure. III.33 Test ANOVA obtenues avec les différentes méthodes pour les fonctions F6, F11 et F22

III.4.2 Deuxième expérience : la classification de différents jeux de données

Dans cette sous-section, la technique VSOS basée sur MLP proposée est utilisée pour classer trois ensembles de données connues [103], les deux premiers sont des bases de données médicales, quant à la dernière, elle concerne un type de fleurs. Les résultats de notre approche sont comparés à ceux obtenus avec d'autres classificateurs ; le SOS original, l'algorithme d'optimisation biographique (BBO : Biogeography-Based Optimizer) [117], l'algorithme génétique (GA) [118], PSO [117] et l'algorithme d'apprentissage incrémentiel basé sur la population (PBIL : Population-Based Incremental Learning) [118]. Nous l'avons également comparé à deux méthodes récentes, la première est un algorithme hybride loup-chauve-souris (HWBA : Hybrid Wolf-Bat Algorithm) pour l'apprentissage de MLP [119] et la seconde est la spécification initiale de SOS basée sur un processus de clustering [33].

III.4.2.1 Bases de données utilisées pour le datamining

Les jeux de données utilisés concernent le cancer du sein et les maladies cardiaquesqui ont été bien détaillés dans le deuxieme chapitre mais la base de données Iris est explicitée ci-dessous [103].

- ***Base de données Iris***

La base de données contient 150 instances, 4 caractéristiques et 3 classes de 50 instances pour chacune, chaque classe fait référence à un type de plante d'iris : Iris Setose, Iris Versicolour et Iris Virginica. Les grands ensembles d'apprentissage donnent une meilleure classification, et les grands ensembles de test donnent une grande confiance dans les métriques évaluées.

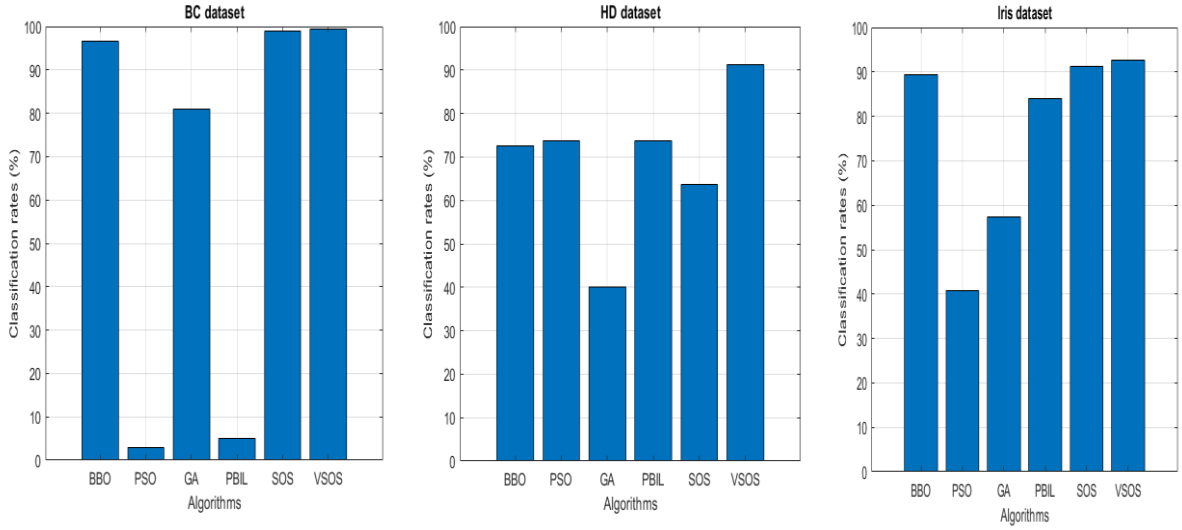
Les caractéristiques des jeux de données utilisés sont résumées dans le tableau III.22, comme indiqué dans la sous-section **III.3.1.**

Tableau. III.22 : Les caractéristiques des bases de données

Base de données	Instances (N)	Caractéristiques	Classe	Nombre d'entraînement	Nombre de teste	Structure de MLP
Cancer du sein (BC)	699	9	2	500	200	**9-19-1**
Cardiaques (HD)	320	22	2	80	320	**22-45-1**
Iris	150	4	3	150	150	**4-9-3**

III.4.2.2 Résultats de la classification

Dans cette expérience, chaque ensemble de données est divisé en ensembles d'entraînement et de test, notre approche MLP proposée basée sur l'algorithme VSOS est utilisée pour l'entraînement et pour trouver les valeurs optimales des poids et des biais. Les résultats de classification obtenus par notre approche, en termes de taux de classification, définis par l'Eq.41, sont représentés sous forme d'histogrammes sur la figure III.34.



a b c

Figure. III.34. Histogrammes des taux de classification : a. Ensemble de données de la BC, b. Jeu de données HD, c. Jeu de données Iris.

D'après la figure III.34, il est clair que notre méthode de classification a obtenu une précision de classification élevée pour tous les ensembles de données utilisés, surpassant toutes les méthodes citées. Dans l'ensemble de données du cancer du sein (BC), la précision de classification de MLP-VSOS est de 99,5 %. Pour les ensembles de données HD (cardiaque) et Iris, MLP-VSOS donne des résultats acceptables ; qui sont respectivement de 91,3% et 92,7%.

Selon ces résultats, notre méthode VSOS peut entraîner avec succès le réseau de neurones MLP et procurer un taux de classification élevé pour différents ensembles de données, ce qui prouve sa robustesse et son efficacité.

Afin de montrer le taux de convergence du MLP-VSOS proposé, la figure III.35 illustre l'évolution de la courbe de fonction de fitness en fonction du nombre d'itérations, obtenu par tous les algorithmes cités, et pour les trois bases de données précédemment décrites. Comme mentionné précédemment, l'évaluation du poids de connexion des neurones est réalisée conformément à la mesure décroissante de l'erreur quadratique moyenne (MSE) en utilisant les sorties du réseau et les sorties observées.

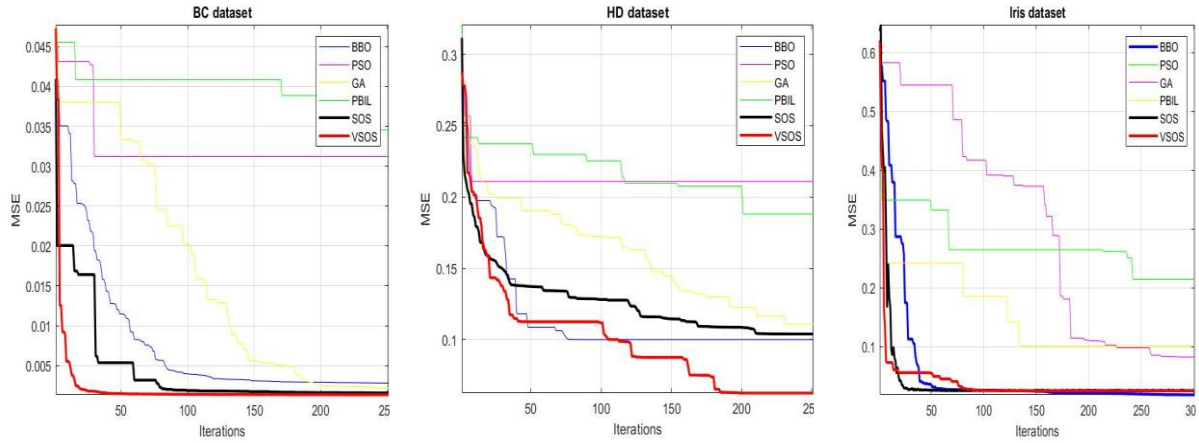


Figure. III.35 Courbes de fonction fitness pour les six algorithmes de classification : a. Ensemble de données de la BC, b. Jeu de données HD, c. Jeu de données Iris.

Comme on peut le voir sur la figure III.35, la courbe MLP-VSOS distinguée par la couleur rouge, a des valeurs décroissantes de la fonction de fitness avec quelques fluctuations. Cela montre que notre MLP-VSOS proposé peut procurer un taux de convergence suffisant, et assurer également une bonne exploration de l'espace de recherche.

Pour montrer la qualité de l'approche de classification proposée et confirmer sa crédibilité, la matrice de confusion (MC) est présentée pour chaque ensemble de données, dans les figures III.36, III.37etIII.38. La spécification de chaque valeur dans cette matrice est définie dans le tableau III.15 de la sous-section III.3.5.

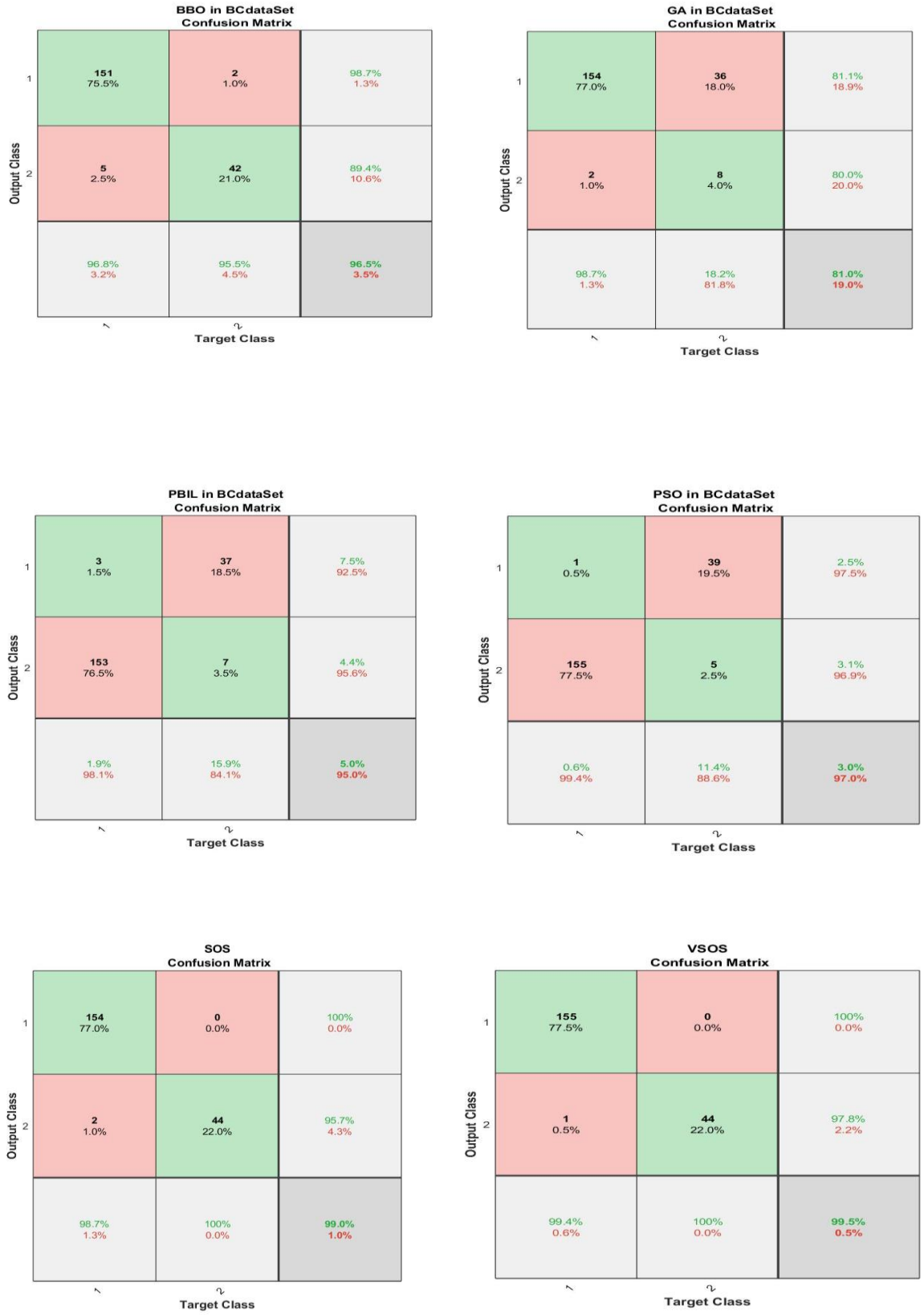


Figure. III.36 Matrices de confusion pour six méthodes appliquées sur la base de données BC

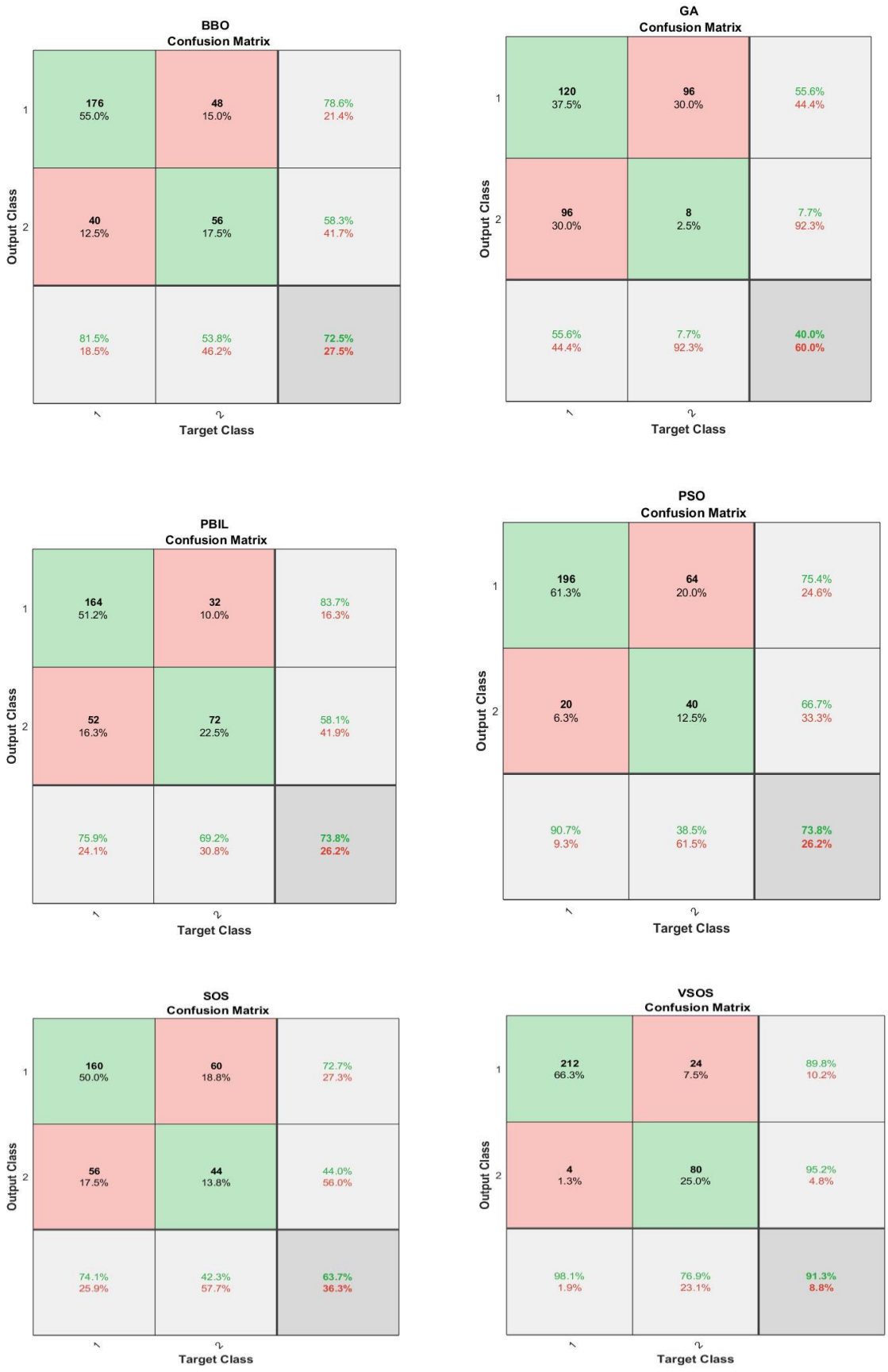


Figure. III.37 Matrices de confusion pour six méthodes appliquées sur la base de données HD

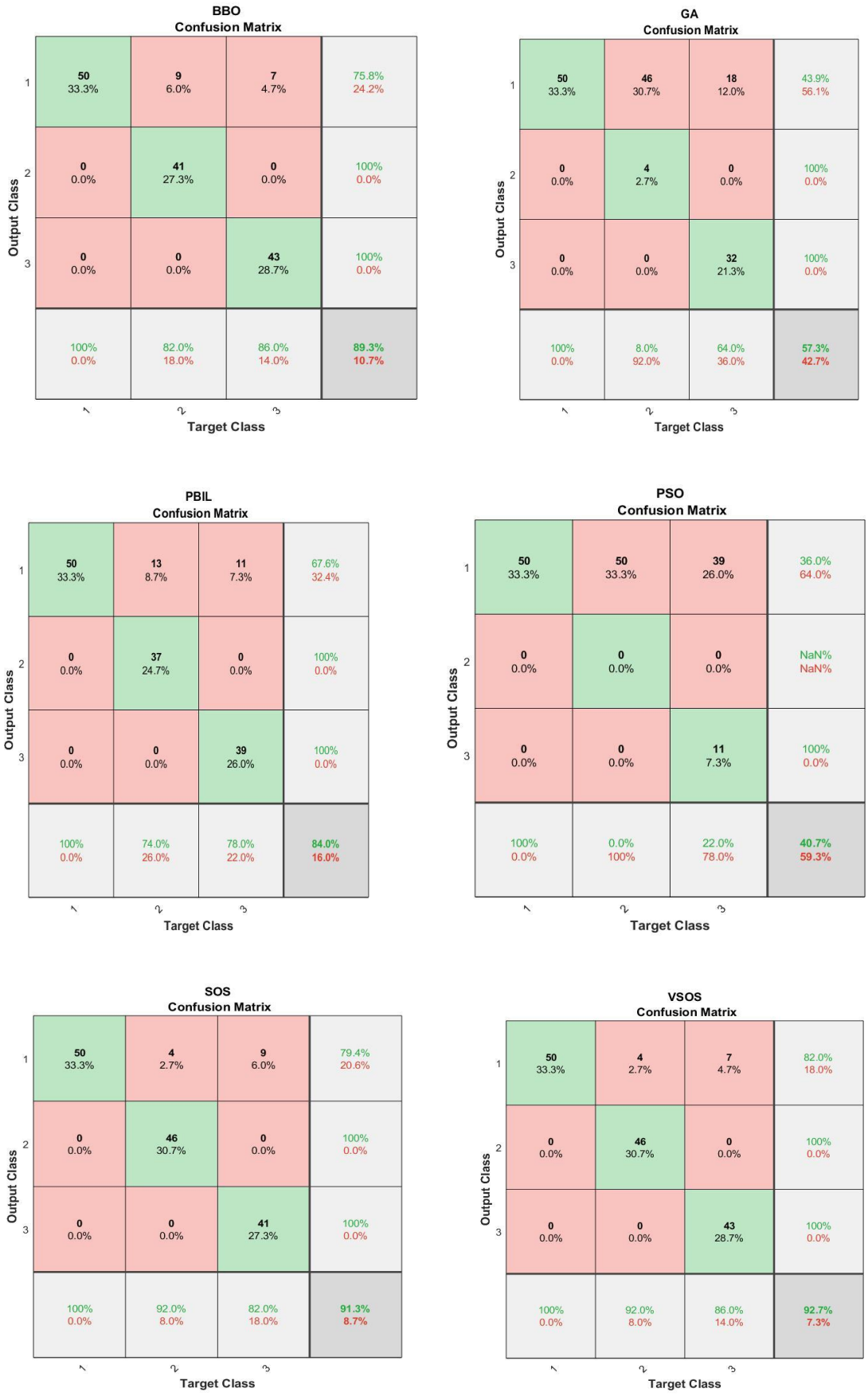


Figure. III.38 Matrices de confusion pour six méthodes appliquées sur la base de données Iris

Les figures III.36, III.37 et III.38 présentent la matrice de confusion obtenue avec les différentes méthodes de classification, elles peuvent fournir une analyse détaillée des résultats de la classification, et contrairement aux autres métriques, MC peut également illustrer si l'ensemble de données est équilibré ou non. La diagonale de la matrice de confusion contient toutes les classifications corrigées.

Dans la base de données BC, sur les 200 échantillons utilisés pour le test, il y en avait 156 dans la première classe et 44 dans la deuxième classe. Notre méthode MLP-VSOS a prédit qu'il y avait 155 échantillons dans la première classe, et 44 dans la seconde, elle n'a qu'une instance mal classée dans la seconde, donc l'éxactitudeest de 99,5 %. Cependant, la méthode MLP-SOS a prédit qu'il y en avait 154 dans la première classe, ce qui signifie que deux instances ont été mal classées, car l'éxactitudeest de 99%.

De même pour les autres méthodes, il est clair que les méthodes citées n'ont pas pu obtenir la prédiction correcte pour tous les échantillons. Selon ces résultats, le MLP-VSOS proposé peut fournir des classificationsmeilleures et dépassant les autres métaheuristiques citées.

Pour plus de précisions sur la qualité de la classification, nous rapportons dans les tableaux III.23, III.24 et III.25les différentes métriques extraites des matrices de confusion pour chaque méthode dans tous les ensembles de données.

Tableau. III.23 Les critères des mesures d'ensemble de données cancer du sein (BC)

Critères (%)	BBO	GA	PBIL	PSO	SOS	HWBA	SOS-Clustering	VSOS
Exactitude	96.5	81.0	05.0	03.0	99.00	96.31	96.48	**99.5**
Rappeler	96.8	98.7	01.9	0.0	100	95.83	99.5	**100**
Sensibilité	98.7	81.8	7.5	0.0	98.7	97.62	95.7	**99.4**
F-mesure	97.74	89.5	3.04	0.0	99.34	96.71	97.56	**99.7**

Tableau. III.24 Les critères des mesures d'ensemble de données sur les maladies cardiaques (HD)

Critères (%)	BBO	GA	PBIL	PSO	SOS	HWBA	SOS-Clustering	VSOS
Exactitude	72.5	40	73.8	73.8	63.7	88.44	65.25	**91.3**
Rappeler	81.5	55.6	75.9	90.7	74.1	81.37	78.9	**98.1**
Sensibilité	78.6	55.6	83.7	75.4	72.7	87.94	74.6	**89.8**
F-mesure	80.02	55.6	79.61	82.35	73.4	84.59	76.68	**93.04**

Tableau. III.25 Métriques de l'ensemble de données Iris

Critères (%)		**BBO**	**GA**	**PBIL**	**PSO**	**SOS**	**VSOS**
Exactitude		89.3	57.3	84	40.7	91.3	**92.7**
Rappeler	Class 1	100	100	100	100	100	**100**
	Class 2	82	8	74	0	92	**92**
	Class3	86	64	48	22	82	**86**
Sensibilité	Class 1	75.8	43.9	67.6	36	79.4	**82**
	Class 2	100	100	100	NaN	100	**100**
	Class3	100	100	100	100	100	**100**
F-measure	Class 1	82.23	61.01	80.7	53	88.52	**90.1**
	Class 2	90.1	14.81	85.06	0	95.83	**95.83**
	Class3	92.47	78.05	64.87	36.1	90.1	**92.47**

Les tableaux III.23, III.24 et III.25 indiquent que la méthode de classification proposée MLP-VSOS peut procurer des résultats très compétitifs ; elle a mieux fonctionné que les autres méthodes citées pour les différents ensembles de données.

Dans l'ensemble de données médicales cancer du sein, MLP-VSOS obtient les meilleurs scores dans tous les critères utilisés (précision, rappel, sensibilité et F-mesure). Notons que l'obtention de résultats élevés avec des bases de données médicales est nécessaire au diagnostic médical. Les autres méthodes citées : SOS-Clustering, SOS original, BBO et le HWBA, elles ont obtenu des résultats acceptables. Pour l'éxactitude, SOS-Clustering avait 96,48%, SOS a obtenu 99%, BBO 96,5% et HWBA 96,31%, et pour le critère de rappel ils ont atteint 99,4%, 96,8% et 95,83% respectivement.

Pour l'ensemble de données HD, VSOS dépasse les autres méthodes ; il a obtenu 91,3 % de précision et 98,1 % de rappel, ce qui signifie que la plupart des personnes diagnostiquées ont été classées dans la bonne classe. Pour l'ensemble de données Iris, notre approche MLP-VSOS a atteint un taux de précision allant jusqu'à 92,7%.

Nous remarquons, dans l'ensemble de données Iris, que le taux de rappel et de précision peut donner des résultats trompeurs ; bien que la précision ne soit pas très élevée, le rappel est de 100 % en classe 1 pour toutes les méthodes, la précision est également de 100 % en classe 2, pour toutes les méthodes, à l'exception de la méthode PSO. Donc, c'est de préférence, de se reposer sur le taux de F-mesure comme qualité de mesure, c'est plus fiable. Ses résultats sont respectivement de 90,1 %, 95,83 % et 92,47 % en classe 1, 2 et 3 pour la méthode VSOS, et ils sont supérieurs à ceux des méthodes citées.

Les résultats présentés ont montré la haute performance de notre classificateur MLP-VSOS et ont confirmé la capacité de notre méthode VSOS à entraîner le réseau MLP en termes de toutes les métriques d'évaluation des performances par rapport aux autres classificateurs

utilisés. A travers l'analyse des résultats expérimentaux, nous pouvons affirmer que notre approche donne une très forte confiance aux experts médicaux dans la prise de décision.

III.5 Conclusion

Dans ce chapitre nous avons en premier lieu décrit une nouvelle version modifiée de l'algorithme SOS appelée VSOS. Un nouveau terme de vitesse est ajouté à l'équation de mise à jourafin de résoudre le problème de convergence prématurée de la méthode SOS conventionnelle. Par la suite, nous avons décrit les fondements théoriques d'une combinaison entre VSOS avec ANN de type MLP.

Nous avons d'autre part présenté aussi les aspects de complexité de l'algorithme proposé pour avoir une estimation qualitative de convergence ce qui nous a conduit à synthétiser que VSOS est très adaptée aux problèmes de classification manipulant des volumes de données assez large ou aussi étroit.

Dans le chapitre suivant nous mettrons en pratique sous forme d'une application de télémédecine à base de VSOS à l'aide à la décision pour le cancer du sein.

Chapitre IV

Troisième contribution :

Prédiction intelligente du cancer du sein à l'aide d'un processus d'exploration de données

IV.1 Introduction

Le cancer du sein est une maladie qui touche des milliers de femmes dans le monde [120]. Il existe actuellement plusieurs méthodes de détection du cancer du sein, telles que : la biopsie, la mammographie et l'échographie (OMS : Organisation Mondiale de la Santé) [121]. Lorsque la tumeur est détectée, elle peut être classée en deux types, le premier est bénin, lorsqu'il n'y a pas de risque de décès, le second est malin lorsqu'il existe un risque de décès (OMS) [121]. La détection de cette maladie au stade initial est très importante, puisqu'un traitement est initié augmentant ainsi le taux de survie. Dans ce chapitre, nous discutons de l'élaboration d'une application à l'aide d'un modèle d'apprentissage automatique formé à l'aide d'une application de prédiction intelligente. Cette application est capable de prédire si la tumeur d'un patient donné est classée comme bénigne ou maligne.

Dans ce chapitre, la nouvelle métaheuristique VSOS proposé dans le chapitre précédent est utilisée dans l'exploration de données pour une étude prédictive du cancer du sein. Nous développerons également notre application intelligente à l'aide de l'algorithme VSOS. Nos expériences ont généré des résultats très confiants en termes de décision prédictive suite à la classification des données.

Ce chapitre est organisé comme suit : dans la section deux dédiée travaux connexes, nous présentons quelques travaux qui ont utilisé le datamining pour le cancer du sein. Ensuite, la troisième section est consacrée à la conception d'une approche intelligente médicale suivie de la section quatre qui contient l'expérimentation et les résultats. La dernière section est destinée à la conclusion et aux perspectives.

IV.2 Rappel de travaux connexes

De nombreux travaux ont été réalisés en utilisant plusieurs méthodes d'exploration de données pour prédire le cancer du sein. Les auteurs [122] ont présenté une approche de prédiction et de diagnostic du cancer du sein à l'aide d'un algorithme de réseaux de neurones. L'auteur [123] a trouvé une méthode pour classer avec précision le cancer du sein comme bénin ou malin à l'aide de Naïve Bayes, Régression Logistique, et d'Arbre de Décision. Cet article [124] a présenté un processus intelligent pour détecter le cancer du sein, en utilisant des réseaux de neurones, et grâce à des données réelles. Les auteurs dans [125] ont proposé le même objectif pour prédire le cancer du sein bénin ou malin, en utilisant deux types de techniques différents, qui étaient les réseaux de neurones convolutifs (CNN) et les réseaux de neurones artificiels multicouches (MLP-ANN) respectivement. L'auteur de [126] a établi un système d'aide à la décision médicale basé sur l'algorithme des arbres J48 dans diverses maladies dont le cancer du sein, qui a donné un bon résultat.

Les auteurs de [127] ont appliqué une stratégie qui comporte deux étapes. La première étape est l'application de l'algorithme des règles d'association, et la seconde est l'utilisation de six classificateurs pour assurer la fiabilité des résultats pour la détection du cancer du sein sur la base de données réelles provenant d'un hôpital iranien. Dans [128] a été proposé un système d'intelligence artificielle (IA) basé sur une méthode d'apprentissage en profondeur, qui peut prédire le cancer du sein grâce au dépistage par mammographie. Ce système, selon ses auteurs, a réduit la charge de travail des experts médicaux de 88 %.

IV.3. Conception et mise en œuvre de l'approche intelligente

Dans cette section, nous proposons l'architecture de l'application intelligente (voir figure IV.39) qui implique la simulation d'un système informatique d'aide à la décision pouvant détecter le cancer à l'aide d'une technique de classification supervisée basée sur l'algorithme VSOS combiné avec ANN de type MLP [129].

L'idée est d'utiliser l'algorithme VSOS pour optimiser les poids et le biais du système ANN en minimisant l'erreur entre les valeurs de sortie réelles et estimées, il représente la fonction de fitness (qui est l'erreur quadratique moyenne (MSE)), qui est réprésentée dans la section III.3.3 par l'équation 41) tâche spécifiée par VSOS. Les principales étapes de ce processus sont présentées comme suit [129] :

1- Prise de données après saisie par le patient ;

2- Le pré-traitement est effectué pour l'optimisation des données, c'est-à-dire le traitement des valeurs manquantes et des données similaires existantes ;

3- Sauvegarder les données sous forme de dossier médical ;

4- Les données sont représentées comme un ensemble de tests par l'approche VSOS-ANN pour la classification supervisée et pour prédire la maladie ;

5- Les résultats obtenus par VSOS à l'étape précédente sont évalués en termes de taux de précision.

L'organigramme de la figure IV.39 ci-dessous montre les détails du processus de candidature :

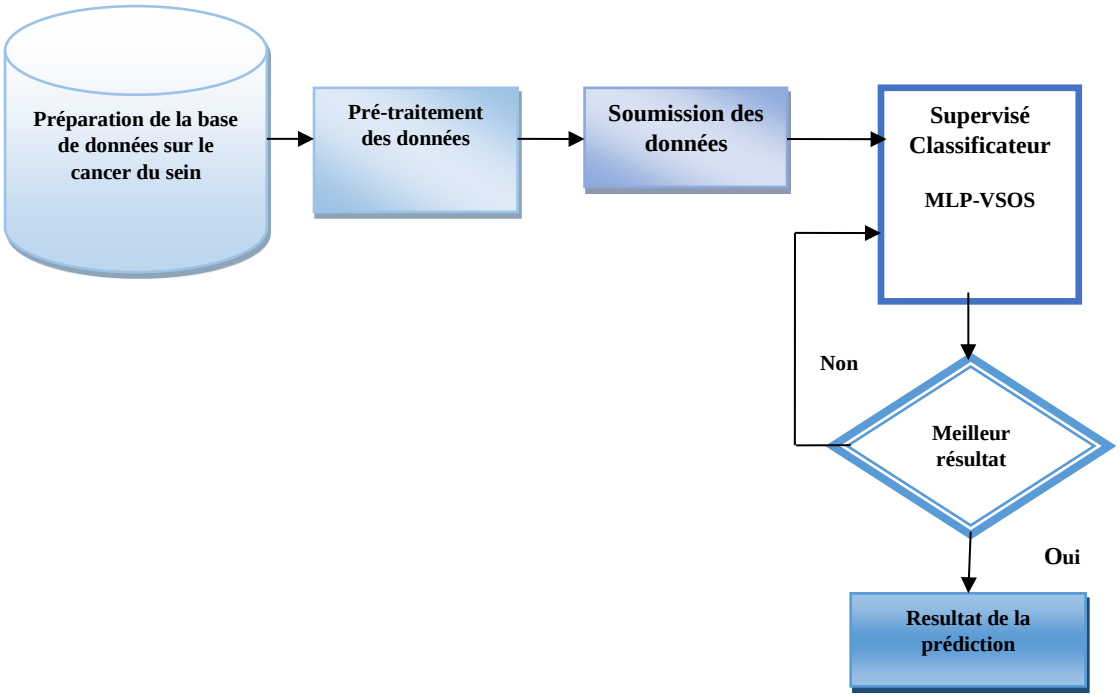


Figure IV.39: Organigramme de l'approche intelligente

• ***Ensemble de données pour l'apprentissage*** *:* pour évaluer notre approche pour le cancer du sein, nous avons utilisé l'ensemble de données sur le cancer du sein de UCI [103]. Cette base contient 699 instances, 9 fonctionnalités et deux classes (où 65,6% pour les tumeurs bénignes et 34,4% pour les tumeurs malignes), nous avons utilisé 500 échantillons comme ensemble d'apprentissage, et l'ensemble de test contenait 200 échantillons. Le modèle a été construit de 19 couches cachées, couches d'entrée avec une couche de sortie (9-19-1). Les caractéristiques de l'ensemble de données sont citées, dans le tableau.II.10 dans le deuxieme chapitre.

Pour donner de la crédibilité à notre méthode, il est possible d'évaluer les résultats expérimentaux en termes de précision, de sensibilité et de spécificité. Ils sont présentés mathématiquement à l'aide du tableau IV.26 :

Tableau. IV.26. Démonstration des formules de critères

Critère	Formule
Exactitude	*= (tp + tn) / (tp + tn + fn + fp)*
Sensitivité	*= (tp)/(tp + fn)*
Specificité	*=(tn)/(tn+fp)*

IV.4. Experimentation

Dans la partie expérimentale, nous avons appliqué l'algorithme VSOS-ANN comme classificateur sur la base de données du cancer du sein.

Tableau IV.27. Résultats des Critères de la Matrice de Confusion

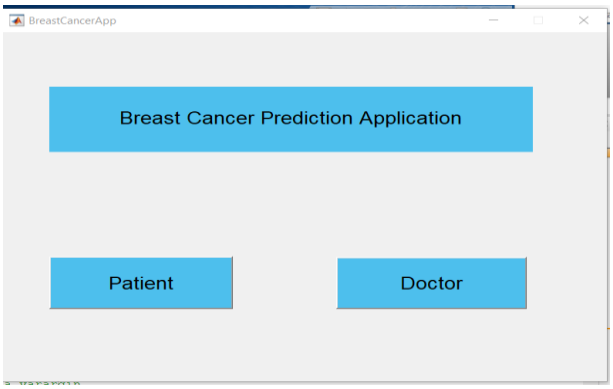
Critère	Resultat (%)
Exactitude	99.50
Sensitivité	100
Specificité	97.8

D'après les résultats obtenus dans le tableau IV.27, avec 100 % de sensibilité, la méthode proposée ne tolère aucune erreur de diagnostic et avec 97,8 % de spécificité elle confirme son efficacité, et fait de la technique VSOS-ANN un bon outil de dépistage.

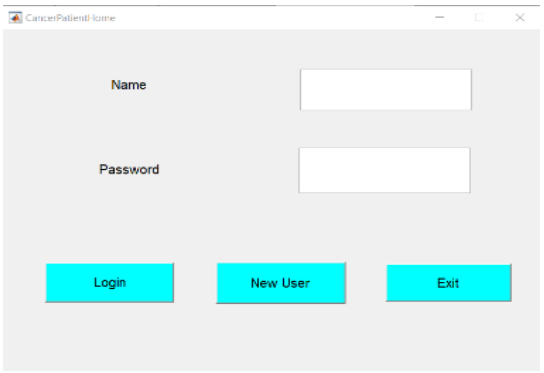
Nous avons également comparé notre travail avec un récent travail [14], il a été développé par Punitha S. et al. basé sur les algorithmes de colonie d'abeilles artificielles (ABC) et de recuit simulé (SA) tous deux hybrides avec ANN, donnant respectivement 99,34 % et 99,11 % comme taux de précision, ces résultats assurent l'avantage de notre technique.

IV.4.1 Fonctionnement de l'approche intelligente

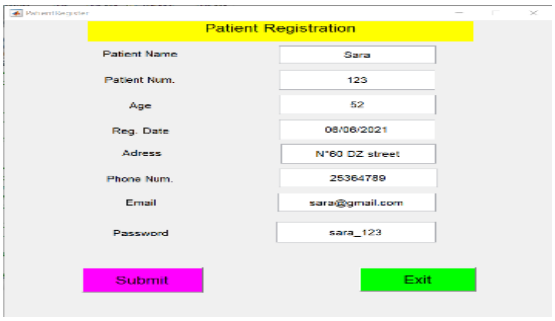
Dans cette section, nous présentons notre application en montrant l'efficacité de VSOS-ANN, pour cette raison nous l'avons exploitée pour déterminer si la patiente avait un cancer du sein bénin ou malin. Par conséquent, nous avons conçu plusieurs interfaces intelligentes qui illustrent clairement notre objectif. Les interactions entre les différents profils d'utilisateurs avec notre application peuvent être réalisées à l'aide des interfaces suivantes :



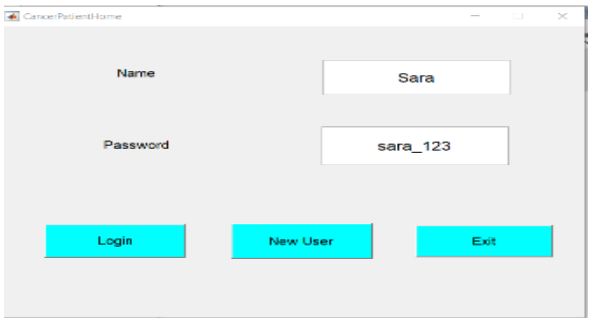
Interface IV.1. Sélectionner l'utilisateur



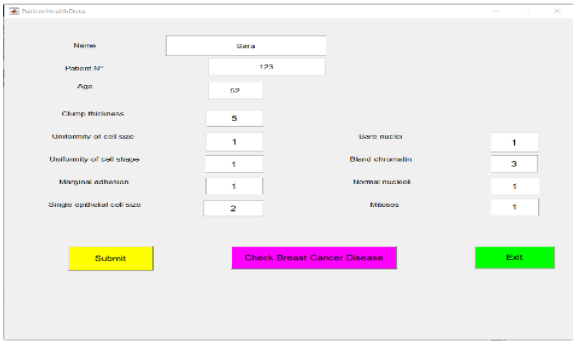
Interface IV.2a. Enregistrement de l'utilisateur



Interface IV.3: Pré-inscription des patients

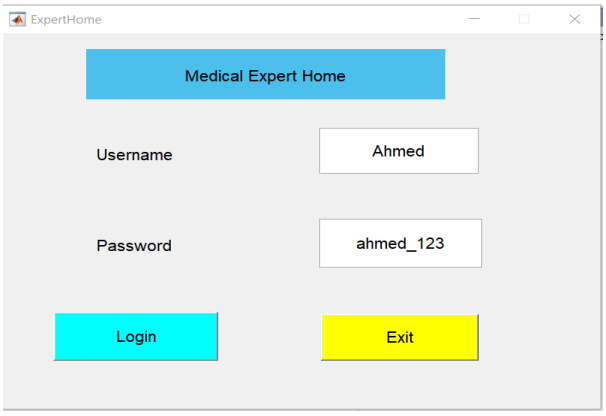


Interface IV..2b.Connexion patient

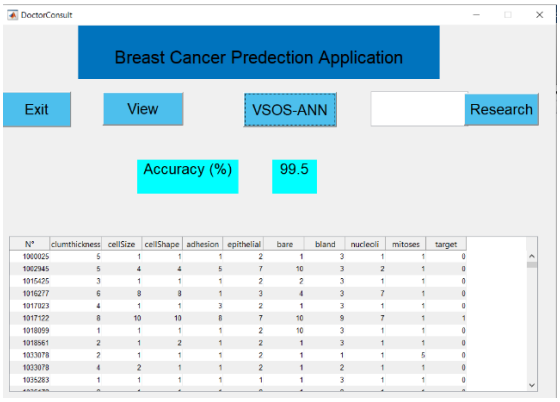


Interface IV.4. Saisie des données et prétraitement

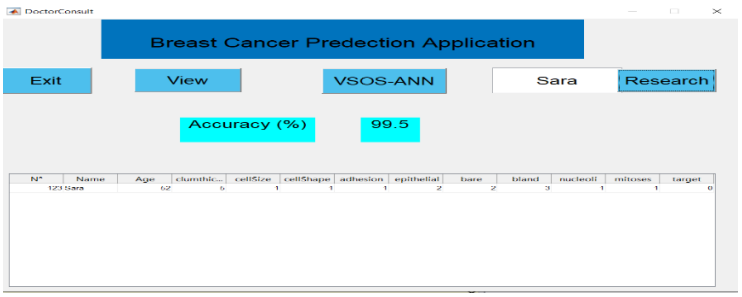
Interface .IV.5. Résultat de la prédiction



Interface IV.6 Connexion expert médical



Interface IV.7a. Résultat de la prédiction de tous les patients



Interface IV.7b. Résultat de la prédiction d'un patient spécifié

IV.4.2. Discussion

La première interface donne les profils des utilisateurs ; il y a deux choix, les patients ou les experts médicaux. Si le choix est le patient, dans la seconde interface une autre sélection doit être effectuée. S'il s'agit d'un cas existant, il se connecte simplement à l'aide de son nom et de son mot de passe, s'il s'agit d'un nouvel utilisateur, il peut s'inscrire via la troisième interface, par exemple un patient nommé ‘‘Sara’’. La quatrième interface récupère les données médicales qui seront envoyées en cliquant sur le bouton ‘‘Submit’’ dans le dossier médical après avoir nettoyé les données inutiles. Un transfert vers la cinquième interface s'effectue en appuyant sur le bouton ‘‘Chek Breast Cancer state’’. Par conséquent, le processus de diagnostic démarre dans l’interface 5 en utilisant le bouton ‘‘Prediction state’’, l'état de prédiction est effectué à l'aide du classificateur de méthode VSOS-ANN. Enfin, le résultat concernant l'état de santé du patient est affiché.

Une fois le processus terminé, l'ensemble des données est transmis en temps réel et répertorié au niveau de l’expert de santé (voir interfaces 6 et 7a). Pour rechercher des donnéesd’un patient spécifique, une option de recherche est proposée en utilisant le nom du patient (voir interface 7b). Par conséquent, l'obtention des informations sur l'état du patient et la décision adéquate concernant le traitement ou le plan de visite sont effectuées.

La figure suivante (figure IV.42) détaille le processus de candidature intelligent.

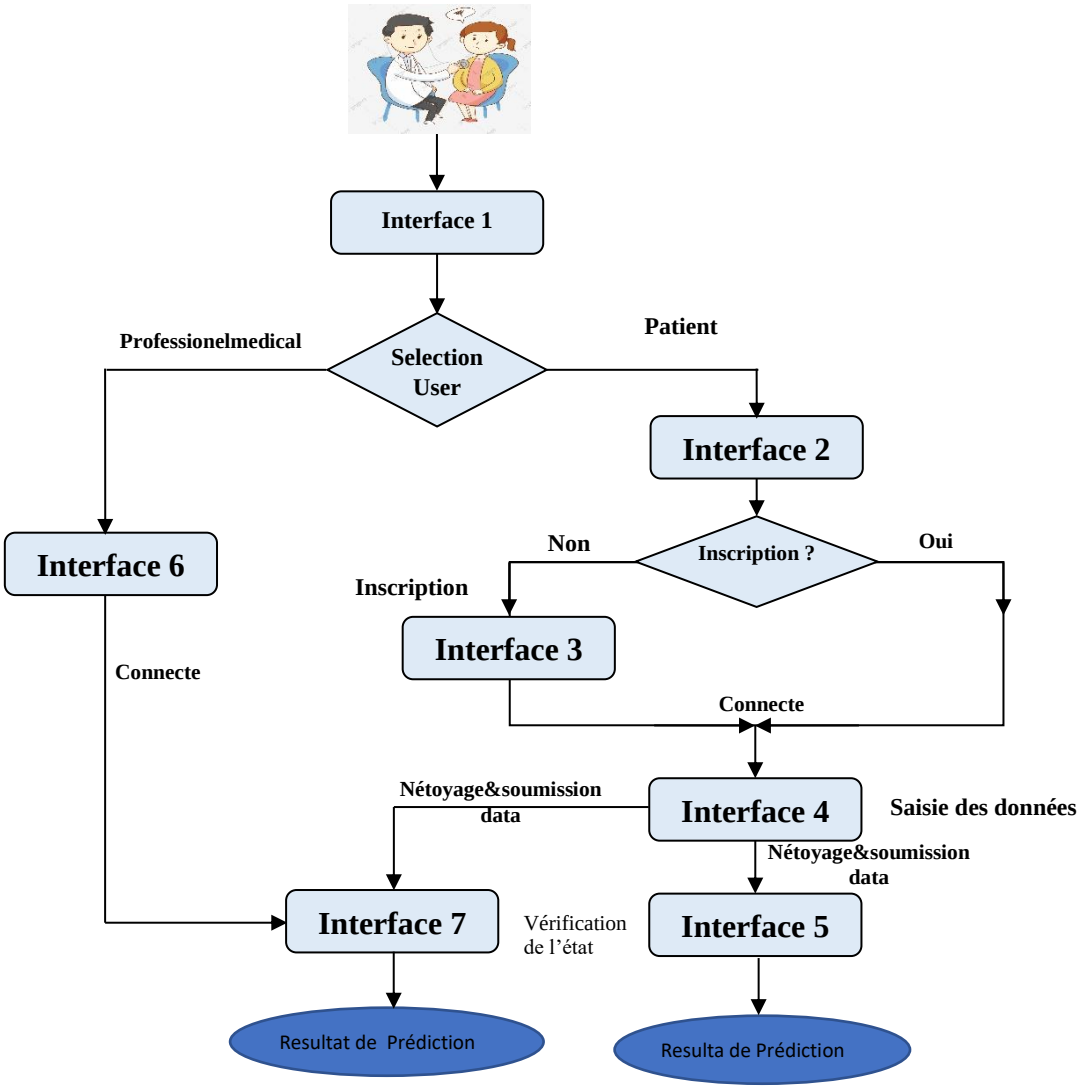


Figure. IV.42 Organigramme des étapes de l'application intelligente

IV.5. Conclusion

L'objectif de ce travail est aidé à la prise de décision sur la nature du cancer qui aide le patient a une intervention plus précoce pour restreindre le développement de la tumeur.

Concernant l'efficacité de la technique VSOS-ANN, nous avons réalisé une application intelligente qui utilise les données médicales du patient et aide à la prise de décision concernant son état de santé. Grâce à ce processus, les professionnels de la santé peuvent réagir rapidement aux interventions nécessaires ou aux modifications des traitements en cours. Le patient peut également obtenir des informations immédiates sur sa situation, ce qui permet d'économiser du temps, des efforts et de l'argent.

CONCLUSION GENERALE ET PERSPECTIVES

CONCLUSION GENERALE ET PERSPECTIVES

Dans le cadre de cette thèse nous avons abordés une problématique d'actualité et de grande importance pour plusieurs domaines de la recherche scientifique et aussi dans les secteurs socioéconomiques manipulant un grand volume de données.

Le traitement qualitatif d'une telle masse de données nécessite le développement et l'exploitation d'approche très rassurante en matière d'analyse de données pour générer une connaissance très précieuse assurant une prise de décision parfois prédictive pour minimiser les risques onéreux notamment quand il s'agit de vie humaine.

C'est ainsi que nous nous sommes intéressés à l'évaluation et l'optimisation d'une solution très prometteuse basée sur l'algorithme SOS à cause de sa simplicité, son efficacité en matière de classification de données et surtout sa convergence rapide.

Pour cela, nous avons d'abord établi un état de l'art exhaustif portant sur le datamining et ses applications et son apport dans plusieurs domaines en particulier dans la médecine.

Nous avons dédié la seconde partie de la présente thèse à nos contributions scientifiques partant de la version initiale de SOS vers son optimisation sous forme de VSOS suivi d'un prototype opérationnel de télémédecine dans le domaine des maladies du cancer du sein.

Nos résultats théoriques et expérimentaux nous ont permis de déduire que l'approche SOS et VSOS qui sont une inspiration d'un phénomène biologique sont des solutions très prometteuse pour le problème de datamining dans le domaine médical.

Certe, il reste plusieurs verrous scientifiques à maitriser et aussi à valider concernant cette approche soit en médecine ou dans d'autres secteurs comme les chaines de production pour la maintenance predictive etc.

C'est ainsi que nos perspectives porterons sur l'exploitation de nos résultats dans le domaine de Big Data qui consistera à valoriser nos contributions sur des volumes de données assez grand et réel telsque ceux des données de COVID-19.

D'autre part, nous envisageons faire un couplage ou une combinaison soit en séquentielle ou en parallèle de SOS et VSOS avec d'autres techniques de datamining (ou Machine Learning) comme les SVM (Support Vector Machine), les RL (Régression Logistique), les DT (Decision Tree).

Une troisième voie mérite aussi une attention dans cet axe de recherche, il s'agit de developper une classification de données basée sur SOS et VSOS associée au Deep Learning au lieu d'un réseau de neurone de type MLP (Perceptron Multi-Couche).

Grâce à cette étude, nous pouvons souligner que plusieurs systèmes de santé en ligne méritent plus de travail de recherche.

Il est très donc très intéressant comme d'autres perspectives d'exploiter la technologie de la blockchain médicale pour partager les données des patients avec l'option de sécurité.

La technologie du cloud computing est également importante pour la e-santé, elle conduit aux services médicaux les plus développés et à leur disponibilité à tout moment et en tout lieu.

Le modèle Bigdata est nécessaire pour la e-santé, le processus de sécurité doit être développé en utilisant une idée intelligente.

Un autre sujet peut être exploré pour la e-santé, il concerne le suivi de l'analyse des sentiments médicaux et l'analyse des données des patients (commentaires, réactions, etc.) sur les réseaux sociaux.

BIBLIOGRAPHIE

BIBLIOGRAPHIE

[1] Sumathi, Sai, and S. N. Sivanandam. Introduction to data mining and its applications. Springer, 2006. vol. 29. p. 1-20.
[2]Teng, Xiuyi et Gong, Yuxia. Research on application of machine learning in data mining. In : IOP conference series: materials science and engineering. IOP Publishing, 2018. p. 062202.doi.org/10.1088/1757-899X/392/6/062202.
[3] Dalvi, Preeti K., Khandge, Siddhi K., Deomore, Ashish, et al. Analysis of customer churn prediction in telecom industry using decision trees and logistic regression. In : 2016 Symposium on Colossal Data Analysis and Networking (CDAN). IEEE, 2016. p. 1-4.
[4] Le, T. T., Watton, J., and Pham, D. T. An artificial neural network-based approach to fault diagnosis and classification of fluid power systems. Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering, 1997, vol. 211, no 4, p. 307-317.
[5] Jason Brownlee. A Gentle Introduction to Bayes Theorem for Machine Learning. on October 4, 2019 in Probability
[6] Pupkov, K. A. et Fadi, I. Collective opinion formation as a set of intelligent agents to achieve the goal. Procedia Computer Science, 2019, vol. 150, p. 216-222.
[7] Abu Alfeilat, Haneen Arafat, Hassanat, Ahmad Ba, Lasassmeh, Omar, et al. Effects of distance measure choice on K-Nearest neighbor classifier performance: A review. Big data, 2019, vol. 7, no 4, p. 221-248.
[8] Hofmann, Heike, Siebes, Arno Pjm, et Wilhelm, Adalbert. Visualizing association rules with interactive mosaic plots. In : Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining. 2000. p. 227-235.
[9] Romero, Cristóbal, Ventura, Sebastián, De Castro, Carlos, et al. Using genetic algorithms for data mining in web-based educational hypermedia systems. In : Proceedings of AH2002 workshop Adaptive Systems for Web-based Education. 2002.
[10] Cheng, Min-Yuan et Prayogo, Doddy. Symbiotic organisms search: a new metaheuristic optimization algorithm. Computers & Structures, 2014, vol. 139, p. 98-112.
[11] Abdullahi, Mohammed, Ngadi, Md Asri, Dishing, Salihu Idi, et al. A survey of symbiotic organisms search algorithms and applications. Neural Computing and Applications, 2019, p. 1-20.
[12] Shawe-Taylor, John et Sun, Shiliang. A review of optimization methodologies in support vector machines. Neurocomputing, 2011, vol. 74, no 17, p. 3609-3618.
[13]NGUYEN, Giang, DLUGOLINSKY, Stefan, BOBÁK, Martin, et al. Machine learning and deep learning frameworks and libraries for large-scale data mining: a survey. Artificial Intelligence Review, 2019, vol. 52, no 1, p. 77-124.
[14] Salem, Osman, Guerassimov, Alexey, Mehaoua, Ahmed, et al. Anomaly detection in medical wireless sensor networks using SVM and linear regression models. International Journal of E-Health and Medical Communications (IJEHMC), 2014, vol. 5, no 1, p. 20-45.
[15] Senthilkumar, D. et Paulraj, S. Prediction of low birth weight infants and its risk factors using data mining techniques. In : Proceedings of the 2015 international conference on industrial engineering and operations management. 2015. p. 186-194.
[16] Duncan, Ian, Loginov, Michael, et Ludkovski, Michael. Testing alternative regression frameworks for predictive modeling of health care costs. North American Actuarial Journal, 2016, vol. 20, no 1, p. 65-87.
[17] Zamuda Aleš, Zarges Christine, Stiglic Gregor, et al. Stability selection using a genetic algorithm and logistic linear regression on healthcare records. In: Proceedings of the Genetic and Evolutionary Computation Conference Companion. 2017. p. 143-144.
[18] Manogaran Gunasekaran et Lopez Daphne. Health data analytics using scalable logistic regression with stochastic gradient descent. International Journal of Advanced Intelligence Paradigms, 2018, vol. 10, no 1-2, p. 118-132.

[19] Govender Prinolan et Ezugwu Absalom E. A symbiotic organisms search algorithm for blood assignment problem. In International workshop on hybrid metaheuristics. Springer, Cham, 2019. p. 200-208.
[20] AYYOUBZADEH, Seyed Mohammad, AYYOUBZADEH, Seyed Mehdi, ZAHEDI, Hoda, et al. Predicting COVID-19 incidence through analysis of google trends data in iran: data mining and deep learning pilot study. JMIR Public Health and Surveillance, 2020, vol. 6, no 2, p. e18828.
[21] Thenmozhi, K. et Deepika, P. Heart disease prediction using classification with different decision tree techniques. International Journal of Engineering Research and General Science, 2014, vol. 2, no 6, p. 6-11.
[22] Iyer Aiswarya, Jeyalatha S., et Sumbaly Ronak. Diagnosis of diabetesusing classification mining techniques. arXivpreprint arXiv:1502.03774, 2015.
[23] Affonso Carlos Sassi, Renato Jose, et Barreiros Ricardo Marques. Biological image classification using rough-fuzzyartificial neural network. Expert Systems with Applications, 2015, vol. 42, no 24, p. 9482-9488.
[24] Mandal Subrata Kumar. Performance analysis of data mining algorithms for breast cancer cell detection using Naïve Bayes, logistic regression and decision tree. International Journal Of Engineering And Computer Science, 2017, vol. 6, no 2.
[25] Samia Noureddine, AbidaToumi, Abir Betka. " Symbiotic Approach for Datamining", ICCES'18 the 4th ACM International Conference of Computing for Engineering and Sciences, Kuala Lumpur, Malaysia — July 06 - 08, (2018), ISBN: 978-1-4503-6447-8, doi>10.1145/3213187.3287610
[26] Oliva, Jefferson Tales et Rosa, João Luís Garcia. Classification for EEG report generation and epilepsy detection. Neurocomputing, 2019, vol. 335, p. 81-95.
[27] Lbahri, O. S., Al-Obaidi, Jameel R., Zaidan, A. A., et al. Helping doctors hasten COVID-19 treatment: Towards a rescue framework for the transfusion of best convalescent plasma to the most critical patients based on biological requirements via ml and novel MCDM methods. Computer methods and programs in biomedicine, 2020, p. 105617.
[28] Mahoto, Naeem Ahmed, Shaikh, Faisal Karim, et Ansari, Abdul Qadir. Exploitation of clustering techniques in transactional healthcare data. Mehran University Research Journal of Engineering & Technology, 2014, vol. 33, no 1, p. 77-92.
[29] Abdel-Maksoud, Eman, Elmogy, Mohammed, et Al-awadi, Rashid. Brain tumor segmentation based on a hybrid clustering technique. Egyptian Informatics Journal, 2015, vol. 16, no 1, p. 71-81.
[30] Sharmili, Kasturi Chandrahaasan et Chilambuchelvan, Arul Gnanaprakasaam. Optimal feature subset selection in high dimensional data clustering. International Journal of Business Intelligence and Data Mining, 2016, vol. 11, no 3, p. 242-263.
[31] Ogbuabor, Godwin et Ugwoke, F. N. Clustering algorithm for a healthcare dataset using silhouette score value. International Journal of Computer Science & Information Technology, 2018, vol. 10, no 2, p. 27-37.
[32] Sohail, Muhammad Noman, Ren, Jiadong, et Uba Muhammad, Musa. A euclidean group assessment on semi-supervised clustering for healthcare clinical implications based on real-life data. International journal of environmental research and public health, 2019, vol. 16, no 9, p. 1581.
[33] Noureddine Samia, Zineeddine Baarir, Toumi Abida, et al. A new predictive medical approach based on data mining and Symbiotic Organisms Search algorithm. International Journal of Computers and Applications, 2020, p. 1-15.
[34] Shalaby, Mohamed, Belal, Nahla A., et Omar, Yasser. Data Clustering Improves Siamese Neural Networks Classification of Parkinson's Disease. Complexity, 2021, vol. 2021.
[35] Jajroudi, M., Baniasadi, T., Kamkar, L., et al. Prediction of survival in thyroid cancer using data mining technique. Technology in cancer research & treatment, 2014, vol. 13, no 4, p. 353-359.

[36] Joshi, Sujata et Nair, Mydhili K. Prediction of heart disease using classification based data mining techniques. In : Computational Intelligence in Data Mining-Volume 2. Springer, New Delhi, 2015. p. 503-511.
[37] Thapen, Nicholas, Simmie, Donal, Hankin, Chris, et al. DEFENDER: detecting and forecasting epidemics using novel data-analytics for enhanced response. PloS one, 2016, vol. 11, no 5, p. e0155417.
[38] Rahim, N. F., Taib, S. M., et Abidin, A. I. Z. Dengue fatality prediction using data mining. Journal of Fundamental and Applied Sciences, 2017, vol. 9, no 6S, p. 671-683.
[39] Nilashi, Mehrbakhsh, Ahmadi, Hossein, Shahmoradi, Leila, et al. A predictive method for hepatitis disease diagnosis using ensembles of neuro-fuzzy technique. Journal of infection and public health, 2019, vol. 12, no 1, p. 13-20.
[40] Sajeev, Shelda, Maeder, Anthony, Champion, Stephanie, et al. Deep Learning to Improve Heart Disease Risk Prediction. In : Machine Learning and Medical Engineering for Cardiovascular Health and Intravascular Imaging and Computer Assisted Stenting. Springer, Cham, 2019. p. 96-103.
[41] Albahri, A. S., Hamid, Rula A., et al. Role of biological Data Mining and Machine Learning Techniques in Detecting and Diagnosing the Novel Coronavirus (COVID-19): A Systematic Review. Journal of Medical Systems, 2020, vol. 44, no 7.
[42] Ayon Safial Islam, Islam, Md Milon, et HOSSAIN, Md Rahat. Coronary artery heart disease prediction: a comparative study of computational intelligence techniques. IETE Journal of Research, 2020, p. 1-20..
[43] Md. Zabirul Islam, Md. Milon Islam and Amanullah Asraf, "A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images", Elsevier, 2020.
[44] Mohmmadzadeh, Hekmat et Gharehchopogh, FarhadSoleimanian. An efficient binary chaotic symbiotic organisms search algorithm approaches for feature selection problems. The Journal of Supercomputing, 2021, p. 1-43.
[45] Methaila, Aditya, Kansal, Prince, Arya, Himanshu, et al. Early heart disease prediction using data mining techniques. Computer Science & Information Technology Journal, 2014, p. 53-59.
[46] Vijayarani, S., Dhayanand, S., et al. Data mining classification algorithms for kidney disease prediction. International Journal on Cybernetics&Informatics (IJCI), 2015, vol. 4, no 4, p. 13-25.
[47] Wu, Haizhou, Zhou, Yongquan, Luo, Qifang, et al. Training feedforward neural networks using symbiotic organisms search algorithm. Computational intelligence and neuroscience, 2016, vol. 2016.
[48] Baitharu, Tapas Ranjan et Pani, Subhendu Kumar. Analysis of data mining techniques for healthcare decision support system using liver disorder dataset. Procedia Computer Science, 2016, vol. 85, p. 862-870.
[49] Jalali, Seyed Mohammad Jafar, Moro, Sérgio, Mahmoudi, Mohammad Reza, et al. A comparative analysis of classifiers in cancer prediction using multiple data mining techniques. International Journal of Business Intelligence and Systems Engineering, 2017, vol. 1, no 2, p. 166-178.
[50] ALJAWARNEH, Shadi, RADHAKRISHNA, Vangipuram, et REDDY, Gali Suresh. Mantra: a novel imputation measure for disease classification and prediction. In : Proceedings of the first international conference on data science, E-learning and information systems. 2018. p. 1-5.
[51] Ayon, Safial Islam, Islam, Md, et al. Diabetes Prediction: A Deep Learning Approach. International Journal of Information Engineering & Electronic Business, 2019, vol. 11, no 2.
[52] Sornalakshmi, M., Balamurali, S., Venkatesulu, M., et al. Hybrid method for mining rules based on enhanced Apriori algorithm with sequential minimal optimization in healthcare industry. Neural Computing and Applications, 2020, p. 1-14.

[53] Int. J Sup. Chain. Applying Data Mining Tools in Transportation: Data-Driven Supply Chain ViewSanjidaBinte Islam#1, Md. Mamun Habib, International Journal of Supply Chain Management IJSCM, Mgt Vol. 10, No. 1, February 202129, ISSN: 2050-7399, pp 29-37
[54] What Is Telemedicine?, American Telemedicine Association, http://www.americantelemed.org /about-telemedicine/what-is-telemedicine#.U4rl, accessed
[55]https://www.argusdelassurance.com/tech/les-cinq-declinaisons-de-la-Telemedecine.151770.
[56] Gheorghe, Mihaela et Petre, Ruxandra. Integrating data mining techniques into telemedicine systems. Informatica Economica, 2014, vol. 18, no 1, p. 120.
[57] https://www.zionmarketresearch.com/news/global-telehealth-market
[58]https://www.transparencymarketresearch.com/pressrelease/telemedicine-market.htm
[59] https://www.mindbowser.com/best-telemedicine-apps/
[60] https://amwell.com
[61] https://mdlnext.mdlive.com/home
[62] https://www.dialogue.co/en/patient-experience
[63] https://www.doctorondemand.com/
[64] https://www.lemonaidhealth.com/
[65] https://doxy.me/en/
[66] https://endoluxe.com/
[67] https://plushcare.com/
[68] https://www.simplecontacts.com/
[69] https://kinesiometrics.org/
[70] https://www.statnews.com/category/first-opinion/
[71] https://welcome.thrive.health/
[72] https://www.dentulu.com/
[73]https://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/telehealth-a-quarter-trillion-dollar-post-covid-19-reality
[74] https://www.micropole.com/fr-fr/actualites/2020-01-20-ces-2020-sante-connectee-le-top-des- innovations-smart-health
[75] Gheorghe, Mihaela et Petre, Ruxandra. Integrating data mining techniques into telemedicine systems. Informatica Economica, 2014, vol. 18, no 1, p. 120.
[76] LAMBERT, L., DESPATIN, J., DHIF, I., et al. Télétransmission en électroencéphalographie et problématique actuelle. SMART-EEG: une solution innovante. European Research in Telemedicine/La Recherche Européenne en Télémédecine, 2015, vol. 4, no 3, p. 81-86.
[77] Albert, Michael, Aschenbrenner, Doris, Barth, Paul, et al. Extension of a telemedicine framework for analysis of industrial machinery data. IFAC-PapersOnLine, 2016, vol. 49, no 30, p. 314-319. 314–319. IFAC (International Federation of Automatic Control) Hosting by Elsevier Ltd. All rights reserved.
[78] Abukhousa, Eman. Analytics and telehealth emerging technologies: the path forward for smart primary care environment. Archivos de Medicina, 2017, vol. 2, no S1, p. 67.
[79] Massaro, Alessandro, Maritati, Vincenzo, Savino, Nicola, et al. A Study of a health resources management platform integrating neural networks and DSS telemedicine for homecare assistance. Information, 2018, vol. 9, no 7, p. 176.
[80] Mlakar, Miha, Puddu, Paolo Emilio, Somrak, Maja, et al. Mining telemonitored physiological data and patient-reported outcomes of congestive heart failure patients. PloS one, 2018, vol. 13, no 3, p. e0190323.
[81] Alelyani, Salem et Ibrahim, Abdelrahman. Internet-of-things in telemedicine for diabetes management. In : 2018 15th learning and technology conference (L&T). IEEE, 2018. p. 20-23.
[82] Anusuya, T. K. et Maharajothi, P. "A Survey of Telemedicine Services using Datamining", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 5 Issue 1, pp. 319-321, January-

February 2019.Available at doi : https://doi.org/10.32628/CSEIT1183818 Journal URL : http://ijsrcseit.com/CSEIT1183818
[83] Vitabile, Salvatore, Marks, Michal, Stojanovic, Dragan, et al. Medical data processing and analysis for remote health and activities monitoring. In : High-Performance Modelling and Simulation for Big Data Applications. Springer, Cham, 2019. p. 186-220.
[84] Ivanov, Borislav et Kralev, Jordan. Re-structuring, Quality Assessment and Telemedicine Integration of an Extracorporeal Perfusion Apparatus Controller. In : 2019 8th Mediterranean Conference on Embedded Computing (MECO). IEEE, 2019. p. 1-4. https://doi.org/10.1007/978-3-030-16272-6_7
[85] Gu, Dongxiao, LI, Tongtong, Wang, Xiaoyu, et al. Visualizing the intellectual structure and evolution of electronic health and telemedicine research. International Journal of Medical Informatics, 2019, vol. 130, p. 103947.doi:10.1016/j.ijmedinf.2019.08.007
[86] Lapointe, Luc, Lavallee-bourget, Marie-Helene, Pichard-jolicoeur, Alexia, et al. Impact of telemedicine on diagnosis, clinical management and outcomes in rural trauma patients: A rapid review. Canadian Journal of Rural Medicine, 2020, vol. 25, no 1, p. 31.
[87] Bernstein, Maria R., John, Libiny, Sciortino, Susan, et al. Does telehealth improve anticoagulation management in patient service centers (PSC)? A pilot project. Journal of Thrombosis and Thrombolysis, 2020, vol. 49, no 2, p. 316-320.. https://doi.org/10.1007/s11239-019-02031-4
[88] Angelucci, Alessandra et Aliverti, Andrea. Telemonitoring systems for respiratory patients: technological aspects. Pulmonology, 2020.. doi:10.1016/j.pulmoe.2019.11.006
[89] Judd, Thomas M., Sloane, Elliot B., et Silva, Ricardo J. The digital hospital of the 21th century, and information systems management. In : Clinical Engineering Handbook. Academic Press, 2020. p. 530-542. doi:10.1016/b978-0-12-813467-2.00081-x
[90] Kruse, Clemens Scott, Lee, Kimberly, Watson, Jeress B., et al. Measures of Effectiveness, Efficiency, and Quality of Telemedicine in the Management of Alcohol Abuse, Addiction, and Rehabilitation: Systematic Review. Journal of medical Internet research, 2020, vol. 22, no 1, p. e13252. DOI: 10.2196/13252 PMID: 32012048
[91] Bashshur, Rashid, Doarn, Charles R., Frenk, Julio M., et al. Telemedicine and the COVID-19 Pandemic, Lessons for the Future. 2020.
[92] Hong, Young-Rock, Lawrence, John, Williams JR, Dunc, et al. Population-level interest and telehealth capacity of US hospitals in response to COVID-19: cross-sectional analysis of Google search and national hospital survey data. JMIR Public Health and Surveillance, 2020, vol. 6, no 2, p. e18961.
[93] Massaad, Elie et Cherfan, Patrick. Social Media Data Analytics on Telehealth During the COVID-19 Pandemic. Cureus, 2020, vol. 12, no 4.
[94] BOEHM, Katharina, ZIEWERS, Stefani, BRANDT, Maximilian P., et al. Telemedicine online visits in urology during the COVID-19 pandemic—potential, risk factors, and patients' perspective. EuropeanUrology, 2020.
[96] Gayathri, V., Mona, M. Chanda, Chitra, S. Banu, et al. A survey of data mining techniques on medical diagnosis and research. International Journal of Data Engineering, 2014, vol. 6, no 6, p. 301-310.
[97] Li, Xiaodong, Tang, Ke, Omidvar, Mohammad N., et al. Benchmark functions for the CEC 2013 special session and competition on large-scale global optimization. gene, 2013, vol. 7, no 33, p. 8.
[98] Wang, Ren, Shi, Keke, et Wang, Shuqi. Research on Performance of Artificial Bee Colony Algorithm Based on Benchmark Test Function. In : IOP Conference Series: Materials Science and Engineering. IOP Publishing, 2020. p. 012019.
[99] Dixit, Abhishek, Kumar, Sushil, Pant, Millie, et al. CA-DE: Hybrid algorithm based on cultural algorithm and DE. In : Machine Intelligence and Signal Analysis. Springer, Singapore, 2019. p. 185-196.

[100] Dixit, Abhishek, Kumar, Sushil, Pant, Millie, et al. CA-DE: Hybrid algorithm based on cultural algorithm and DE. In : Machine Intelligence and Signal Analysis. Springer, Singapore, 2019. p. 185-196.
[101] Ma, Zhiteng, Yuan, Xianfeng, Han, Sen, et al. Improved chaotic particle swarm optimization algorithm with more symmetric distribution for numerical function optimization. Symmetry, 2019, vol. 11, no 7, p. 876.
[102] ARora, Sankalap et Singh, Satvir. Butterfly optimization algorithm: a novel approach for global optimization. Soft Computing, 2019, vol. 23, no 3, p. 715-734.
[103] Lichman, M. UCI Machine Learning Repository; University of California, School of Information and Computer Science: Irvine, CA, USA, 2013; Available online: http://archive.ics.uci.edu/ml (accessed on 13 March 2016).
[104] Haq, Ehtasham-ul, Ahmad, Ishfaq, Hussain, Abid, et al. A novel selection approach for genetic algorithms for global optimization of multimodal continuous functions. Computational intelligence and neuroscience, 2019, vol. 2019.
[105] M. Krzywinski and N.Altman. Points of significance: visualizing samples with box plots. Nat Methods, 2014, vol. 11, pp. 119-120.
[106] Visa, Sofia, Ramsay, Brian, Ralescu, Anca L., et al. Confusion matrix-based feature selection. MAICS, 2011, vol. 710, p. 120-127.
[107] Aljarah, Ibrahim, Mafarja, Majdi, Heidari, Ali Asghar, et al. Clustering analysis using a novel locality-informed gray wolf-inspired clustering approach. Knowledge and Information Systems, 2020, vol. 62, no 2, p. 507-539.
[108] R. Kummar and A. Indrawn. Receiver Operating Characteristic (ROC) Curve for Medical Researchers. Indian Pediatrics, 2011, vol. 48, p.277-289.
[109] Cinar, Ahmet Cevahir. Training feed-forward multi-layer perceptron artificial neural networks with a tree-seed algorithm. Arabian Journal for Science and Engineering, 2020, vol. 45, no 12, p. 10915-10938.
[110] Hecht-nielsen, Robert. Kolmogorov's mapping neural network existence theorem. In : Proceedings of the international conference on Neural Networks. New York, NY, USA : IEEE Press, 1987. p. 11-14.
[111] Congress on Evolutionary Computation, Proceedings of the IEEE Congress, CEC 2005, 2-4 September 2005, Edinburgh, UK. IEEE 2005, ISBN 0-7803-9363-5.
[112] Hashim, Fatma A., Hussain, Kashif, Houssein, Essam H.,et al, (2020). Archimedes optimization algorithm: a new metaheuristic algorithm for solving optimization problems. Applied Intelligence. 1-21.
[113] Chou, J. S., & Truong, D. N. (2021). A novel metaheuristic optimizer inspired by behavior of jellyfish in ocean. Applied Mathematics and Computation, 389, 125535.
[114] Betka, Abir., Terki, Nadjiba, Toumi, Abida, & Dahmani, H. (2020). Grey wolf optimizer-based learning automata for solving block matching problem. Signal, image and video processing, 14(2), 285-293.
[115] Tejani, G. G., Savsani, V. J., & Patel, V. K. (2016). Adaptive symbiotic organisms search (SOS) algorithm for structural design optimization. Journal of Computational Design and Engineering, 3(3), 226-249.
[116] Dinh-Cong, D., Nguyen-Thoi, T., & Nguyen, D. T. (2020). A FE model updating technique based on SAP2000-OAPI and enhanced SOS algorithm for damage assessment of full-scale structures. Applied Soft Computing, 89, 106100.
[117] Mirjalili, S., Mirjalili, S. M., & Lewis, A. (2014). Let a biogeography-based optimizer train your multi-layer perceptron. Information Sciences, 269, 188-209.
[118] Ankrah, R., Regnier-Coudert, O., McCall, J., Conway, A., & Hardwick, A. (2018, July). Performance analysis of ga and pbil variants for real-world location-allocation problems. In 2018 IEEE Congress on Evolutionary Computation (CEC) (pp. 1-8). IEEE.

[119] Agrawal, U., Arora, J., Singh, R., Gupta, D., Khanna, A., & Khamparia, A. (2020). Hybrid Wolf-Bat Algorithm for Optimization of Connection Weights in Multi-layer Perceptron. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 16(1s), 1-20.
[120] Desantis, Carol E., Ma, Jiemin, Gaudet, Mia M., et al. Breast cancer statistics, 2019. CA: a cancer journal for clinicians, 2019, vol. 69, no 6, p. 438-451.
[121] Koroltchouk, Valentin, Stanley, Kenneth, et Stjernswärd, Jan. The control of breast cancer a World Health Organization perspective. Cancer, 1990, vol. 65, no 12, p. 2803-2810.
[122] Arya, C., & Tiwari, R. (2016, January). Expert system for breast cancer diagnosis: A survey. In 2016 international conference on computer communication and informatics (ICCCI) (pp. 1-9). IEEE.
[123] Christopher, Jabez J., Nehemiah, Khanna H., et Arputharaj, Kannan. Knowledge-based systems and interestingness measures: Analysis with clinical datasets. Journal of computing and information technology, 2016, vol. 24, no 1, p. 65-78.
[124] Abdullahi, Mohammed, Ngadi, Md Asri, Dishing, Salihu Idi, et al. A survey of symbiotic organisms search algorithms and applications. Neural Computing and Applications, 2019, p. 1-20.
[125] Mandal, Subrata Kumar. Performance analysis of data mining algorithms for breast cancer cell detection using Naïve Bayes, logistic regression and decision tree. International Journal Of Engineering And Computer Science, 2017, vol. 6, no 2.
[126] Marazzi, Fabio, Tagliaferri, Luca, Masiello, Valeria, et al. Generator Breast DataMart—The Novel Breast Cancer Data Discovery System for Research and Monitoring: Preliminary Results and Future Perspectives. Journal of Personalized Medicine, 2021, vol. 11, no 2, p. 65.
[127] Ahmed, Luqman, Iqbal, Muhammad Munwar, Aldabbas, Hamza, et al. Images data practices for semantic segmentation of breast cancer using deep neural network. Journal of Ambient Intelligence and Humanized Computing, 2020, p. 1-17.
[128] Kim, Jaeil, Kim, Hye Jung, Kim, Chanho, et al. Artificial intelligence in breast ultrasonography. Ultrasonography, 2021, vol. 40, no 2, p. 183.
[129] Samia Noureddine, Baarir Zineeddine, Abida Toumi, Abir Betka, Nesrine Kazar, Aicha-Nabila Benharkat, "Smart Breast Cancer Prediction Using Data Mining Process",2021, International Conference on Information Systems and Advanced Technologies (ICISAT). IEEE Xplore: 20 January 2022. ISBN:978-1-6654-7824-3. DOI :10.1109/ICISAT54145.2021.9678429

ANNEXE

ANNEXE A : Les fonctions du teste

Fonctions	D	Interval	Optimum
$f_{01} = \sum_{i=1}^{n} x_i^2$	30	$[-100, +100]$	0
$f_{02} = \sum_{i=1}^{n} \lvert x_i \rvert + \prod_{i=1}^{n} \lvert x_i \rvert$	30	$[-10, +10]$	0
$f_{03} = \sum_{i=1}^{n} \left(\sum_{j=1}^{i} x_j \right)^2$	30	$[-100, +100]$	0
$f_{04} = max_i\{\lvert x_i \rvert, 1 \le i \le D\}$	30	$[-100, +100]$	0
$f_{05} = \sum_{i=1}^{D-1} [100(x_{i+1} - x_i^2)^2 + (x_i - 1)^2]$	30	$[-30, +30]$	0
$f_{06} = \sum_{i=1}^{D} ([x_i + 0.5])^2$	30	$[-100, +100]$	0
$f_{07} = \sum_{i=1}^{D} i x_i^4 + random\,[0,1]$	30	$[-1.28, +1.28]$	0
$f_{08} = \sum_{i=1}^{D} -x_i sin\left(\sqrt{\lvert x_i \rvert}\right)$	30	$[-500, +500]$	-418.9829 * n
$f_{09} = \sum_{i=1}^{D} [x_i^2 - 10\,cos(2\pi x_i) + 10]$	30	$[-5.12, +5.12]$	0
$f_{10} = -20\,exp\left(-0.2\sqrt{\frac{1}{D}\sum_{i=1}^{D} x_i^2}\right) - exp\left(\frac{1}{D}\sum_{i=1}^{D} cos(2\pi x_i)\right) + 20 + e$	30	$[-32, +32]$	8.8818e-16
$f_{11} = \frac{1}{4000}\sum_{i=1}^{D} x_i^2 - \prod_{i=1}^{D} cos\left(\frac{x_i}{\sqrt{i}}\right) + 1$	30	$[-600, +600]$	0
$f_{12} = \frac{\pi}{D}\left\{10\,sin^2(\pi y_i) + \sum_{i=1}^{D-1} (y_i - 1)^2[1 + 10sin^2(\pi y_i + 1)] + (yD - 1)^2 + \sum_{i=1}^{D} u(x_i, 10, 100, 4)\right\}$ $y_i = 1 + \frac{x_i + 1}{4}$, $u(x_i, a, k, m) = \begin{cases} k(x_i - a)^m x_i > a \\ 0 \quad -a < x_i < a \\ k(-x_i - a)^m x_i < -a \end{cases}$	30	$[-50, +50]$	1.5705e-032
$f_{13} = \left[\frac{1}{500} + \sum_{j=1}^{25} \frac{1}{j + \sum_{i-1}^{2}(x_i - a_{ij})^6}\right]^{-1}$	2	$[-65.53, +65.53]$	0.998004
$f_{14} = \sum_{i=1}^{11} \left[a_i - \frac{x_1(b_i^2 + b_i x_i)}{b_i^2 + b_i x_3 + x_4}\right]^2$	4	$[-5, +5]$	0.0003075
$f_{15} = 4x_1^2 - 2.1x_i^4 + \frac{1}{3}x_1^6 + x_1 x_2 - 4x_2^2 + 4x_2^4$	2	$[-5, +5]$	-1.0316285
$f_{16} = \left(x_2 - \frac{5.1}{4\pi^2}x_1^2 + \frac{5}{\pi}x_1 - 6\right)^2 + 10\left(1 - \frac{1}{8\pi}\right)cosx_1 + 10$	2	$[5, +10]$ $* [0, +15]$	0.398
$f_{17} = [1 + (x_1 + x_2 + 1)^2(19 - 14x_1 + 3x_1^2 - 14x_2 + 6x_1x_2 + 3x_2^2)] \times [30 + (2x_1 - 3x_2)^2(18 - 32x_1 + 12x_1^2 + 48x_2 - 36x_1x_2 + 27x_2^2)]$	2	$[-5, +5]$	3
$f_{18} = -\sum_{i=1}^{4} c_i\,exp\left(-\sum_{j=1}^{3} a_{ij}(x_j - p_{ij})^2\right)$	3	$[0, +1]$	-3.86
$f_{19} = -\sum_{i=1}^{4} c_i\,exp\left(-\sum_{j=1}^{6} a_{ij}(x_j - p_{ij})^2\right)$	6	$[0, +1]$	-3.32
$f_{20} = -\sum_{i=1}^{5} [(X - a_i)(X - a_i)^T + c_i]^{-1}$	4	$[0, +10]$	-10.1532
$f_{21} = -\sum_{i=1}^{7} [(X - a_i)(X - a_i)^T + c_i]^{-1}$	4	$[0, +10]$	-10.4029
$f_{22} = -\sum_{i=1}^{10} [(X - a_i)(X - a_i)^T + c_i]^{-1}$	4	$[0, +10]$	-10.5364