

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE
UNIVERSITE MOHAMED BOUDIAF - M'SILA

FACULTE DE TECHNOLOGIE
DEPARTEMENT D'ELECTRONIQUE
N° :



DOMAINE : SCIENCES ET TECHNOLOGIE
FILIERE : GENIE ELECTRONIQUE
OPTION : SYSTEME EMBARQUE

**Mémoire présenté pour l'obtention
Du diplôme de Master Académique**

Par:

BACHIRI Mohamed
DAKHANE Mohammed

Intitulé

**ÉVALUATION ET DETERMINATION DES VARIABLES D'ENTREE POUR UN
MODELE MULTICLASSE INTELLIGENT DANS LES PROCÉDÉS DE
TRAITEMENT DES EAUX PROPRES**

Soutenu devant le jury composé de:

Mr. H. MEZAACHE	Université de M'sila	Président
Dr. M. LADJAL	Université de M'sila	Encadreur
Mr. Y. BRIK	Université de M'sila	Co-Encadreur
Mr. M. DJERIOUI	Université de M'sila	Examineur

Année universitaire : 2017 /2018

Remerciements

Tout d'abord, Nous remercions DIEU miséricorde de nous avoir donné la volonté, le courage et la patience qui nous ont permis de réaliser ce travail, et de poursuivre nos études supérieures et de les réussir et d'avoir la chance d'atteindre le niveau MASTER II.

Nous tenons à exprimer nos vifs remerciements à nos encadreurs Mr. BRIK Youssef et Mr. LADJAL Mohamed d'avoir proposée ce sujet et pour avoir dirigé nos travail durant la période d'élaboration de ce mémoire.

Nous les exprimer notre profonde gratitude pour nous avoir fait profiter de leurs connaissances, mais aussi de leurs méthodes de travail, et surtout de leur rigueur scientifique.

Nos remerciements s'adressent également à tous membres de Jury, qui ont accepté de nous honorer de leur présence et de juger notre travail.

Nos remerciements vont aussi à tous les enseignants du département d'électronique sans oublier toutes les personnes de ce même département.

Enfin, nous remercions A tous ceux qui ont contribué de près ou de loin à la réalisation de ce projet de fin d'études.

Dédicaces

Je dédie ce modeste travail :

À mes très chers parents pour tous leurs sacrifices, leur amour, leur tendresse, leur soutien et leurs prières tout au long de mes études,

Je les remercie de m'avoir encouragé

Et aidés à devenir ce que je suis.

À mes chers frères et chers sœurs, pour leur appui et leur encouragement,

À toute ma famille pour leur soutien tout au long de nos parcours universitaire

À tous mes enseignants.

Et à toute mes très chers amis, mes très chers collègues, et toute la promotion Master De l'année 2018.

Dakhane Mohammed.

Dédicace

Je dédie ce modeste travail :

*A ma mère pour ses sacrifices en témoignage de tout mon
affection au long de mes études.*

*À mon cher père, que j'espère que Dieu préservera avec Sa grande
miséricorde et demeurera dans Son paradis*

A ma chère fiancée AZZA Souhila

A mes chers frères et chers sœurs, pour leur appui et leur encouragement,

*A toute ma famille pour leur soutien tout au long de nos parcours
universitaire*

A tous mes enseignants.

*Et a toute mes très chers amis, mes très chers collègues, et toute la
promotion Master De l'année 2018.*

Bachiri Mohamed.

Table des matières

LISTE DES FIGURES	iv
LISTE DES TABLEAUX	v
ABREVIATION ET NOTATION.....	vi
INTRODUCTION GENERALE	1

CHAPITRE I GENERALITES SUR LES EAUX PROPRES

I.1. INTRODUCTION	4
I.2. GENERALITES DES EAUX	4
I.2.1. Cycle de l'eau	4
I.2.2. Ressources des eaux	5
I.2.2.1. Les eaux de pluie	6
I.2.2.2. Les eaux de surface	6
I.2.2.3. Les eaux souterraines	7
I.2.3. Importance de l'eau	7
I.2.4. le problème d'eaux	8
I.2.5. L'eau potable	8
I.2.6. Normes de la qualité des eaux	9
I.2.7. Importance de l'analyse et du traitement.....	10
I.3. CHAINE ELEMENTAIRE DE PRODUCTION D'EAU POTABLE	11
I.3.1. Prétraitement.....	12
I.3.2. Pré-oxydation	13
I.3.3. Clarification	14
I.3.4. Oxydation-Désinfection	15
I.3.5. Affinage.....	16
I.4. LES CAPTEURS.....	16
I.4.1. Mesure des paramètres physico-chimiques	17
I.4.1.1. Mesure des paramètres usuels	17
I.4.1.2. Mesure des paramètres spécifiques	18
I.4.2. Capteurs physiques.....	18
I.5. METHODES DES SURVILLANCE	20
I.5.1. Méthodes de surveillance avec modèles.....	21
I.5.1.1. Redondances physiques et analytiques.....	21

I.5.1.2. Méthodes d'estimation paramétrique	22
I.5.2. Méthode de surveillance sans modèle	23
I.5.2.1. Surveillance utilisant des outils statistiques de traitement du signal.....	23
I.5.2.2. Surveillance par des méthodes à partir de données historiques	23
I.5.2.3. Les méthodes à base de modèles comportementaux	24
I.5.2.4. Les méthodes de reconnaissance de formes	24
I.5.2.5. Les méthodes à base de modèles explicatifs	24
I.6.SURVEILLANCE PAR RECONNAISSANCE DE FORMES.....	25
I.6.1.Objectif de ce travail	26
I.7.CONCLUSION	27

CHAPITRE II SYSTEME PROPOSE

II.1. INTRODUCTION.....	28
II.2. SYSTEME DE LA RECONNAISSANCE DES FORMES	28
II.2.1. Présentation de la reconnaissance des formes.....	28
II.2.2. Schéma général d'un système de reconnaissance	29
II.2.2.1. Système d'acquisition et prétraitement	30
II.2.2.2. Extraction de caractéristiques	30
II.2.2.3. Classification	31
II.3. RÉDUCTION DE DIMENSION DES DONNÉES.....	32
II.3.1. Les méthodes de Réduction de dimension des données	33
II.3.1.1. Neighborhood Component Analysis (NCA).....	34
II.3.1.2. Algorithme de NCA	36
II.4. CLASSIFICATION	36
II.4.1. Classification non supervisée	37
II.4.2. Classification supervisée	37
II.4.3. Classification par séparation	38
II.4.3.1. Réseaux de neurones	38
II.4.3.1.1. Le neurone formel	38
II.4.3.1.2. Principe de fonctionnement	39
II.4.3.1.3. Fonction d'activation	39
II.4.3.1.4. Topologies de réseaux de neurones	40
II.4.3.1.5. Le Perceptron	41
II.4.3.1.6. Le perceptron multicouche	42

II.4.3.1.7. Algorithmes d'apprentissage des RNAs	43
II.5.CONCLUSION	46

CHAPITRE III ÉVALUATION EXPERIMENTALE

III.1. INTRODUCTION	47
III.2. DEMARCHE PROPOSEE	47
III.3. SYSTEME DE CONTROLE ET DE SURVEILLANCE	48
III.3.1. Approche utilisée dans la surveillance.....	50
III.4. DESCRIPTION DES DONNEES D'ENTREE.....	50
III.4.1. Les Bases de données	51
III.5. PARTIE EXPÉRIMENTALE	51
III.5.1. Evaluation de NCA	52
III.5.2. Evaluation de RNAs	55
III.5.3. L'influence du nombre de neurones dans la couche cachée de RNA.....	57
III.5.4. Le choix de la fonction d'activation pour le RNA	59
III.6. CONCLUSION.....	61
CONCLUSION GENERALE.....	62
REFERENCES	64

LISTE DES FIGURES

Fig. I.1. Cycle des eaux	5
Fig. I.2. Station de production d'eau potable	12
Fig. I.3. Classification des méthodologies de surveillance industrielle	21
Fig. I.4. Schéma général du système de surveillance par reconnaissance de formes	26
Fig. II.1. Schéma général d'un système de reconnaissance de formes.	29
Fig. II.2. Caractéristique extraites sur des formes appartenant a 3classes dans un espace de représentation bidimensionnel $d = 2$ et $K = 3$	31
Fig. II.3. Schéma d'un neurone formel	39
Fig. II.4. Les différentes formes de la fonction d'activation.....	40
Fig. II.5. Exemple de réseau multicouche a 1 couche cachée, 3 entrées et deux sorties....	40
Fig. II.6. Exemple de réseau a connexions récurrentes.	41
Fig. II.7. Exemple de réseau a connexions complètes.	41
Fig. II.8. Représentation graphique de perceptron	42
Fig. II.9. Architecture général d'un réseau de neurone MLP.....	43
Fig. II.10. Exemple de réseau MLP a une couche cachée avec 5 entrées ,3 neurones dans la couche cachée, et quatre sorties	44
Fig. III.1. Architecture du système de contrôle et de surveillance.	49
Fig. III.2. Poids de chaque caractéristique en utilisant NCA sur la base 4D	52
Fig. III.3. poids de chaque caractéristique en utilisant NCA sur la base 8D	54
Fig. III.4. La correspondance entre la sortie désirée et de la sortie de RNA dans le cas d'apprentissage et de test sur la base 4D.	56
Fig. III.5. La correspondance entre la sortie désirée et de la sortie de RNA dans le cas d'apprentissage et de test sur la base 8D.	57
Fig. III.6. Taux de classification (test) en fonction de nombre de neurones dans la couche cachée sur la base 4D.	58
Fig. III.7. Courbe montre le taux de reconnaissance de l'RNA en fonction de nombre de neurones dans la couche cachée sur la base 4D.....	58

LISTE DES TABLEAUX

Tab. I.1. Estimation de la distribution globale de l'eau	6
Tab. I.2. Normes d'eau potable selon l'OMS (2006).	9
Tab. I.3. Classification des eaux d'après leur pH.....	19
Tab. III.1. Résultats d'apprentissage sur la base 4D en utilisant K-fold-cross-validation avec 4 combinaisons des descripteurs	53
Tab. III.2. Résultats de test sur la base 4D en utilisant K-fold-cross-validation avec 4 D combinaisons des descripteurs.....	53
Tab. III.3. Résultats d'apprentissage sur la base 8D en utilisant K-fold-cross-validation avec 8 combinaisons des descripteurs.	54
Tab. III.4. Résultats de test sur la base 8D en utilisant K-fold-cross-validation avec 8 D combinaisons des descripteurs.....	54
Tab. III.5. Résultats d'apprentissage de RNA sur la base 4D en utilisant K-fold-cross-validation avec différentes combinaisons des descripteurs..	55
Tab. III.6. Résultats de test de RNA sur la base 4D en utilisant K-fold-cross-validation avec différentes combinaisons des descripteurs..	55
Tab. III.7. Résultats d'apprentissage de RNA sur la base 8D en utilisant K-fold-cross-validation avec différentes combinaisons des descripteurs..	56
Tab. III.8. Résultats de test de RNA sur la base 8D en utilisant K-fold-cross-validation avec différentes combinaisons des descripteurs..	56
Tab. III.9. Taux d'apprentissage selon la fonction d'activation de RNA sur la base 4D	59
Tab. III.10. Taux de test selon la fonction d'activation de RNA sur la base 4D	59
Tab. III.11. Taux d'apprentissage selon la fonction d'activation de RNA sur la base 8D	60
Tab. III.12. Taux de test selon la fonction d'activation de RNA sur la base 8D	60

ABREVIATIONS ET NOTATIONS

IA : Intelligence Artificielle
RNA : Réseaux de Neurones Artificiels
NCA : Neighborhood Component Analysis
SVM : Support Vector Machines
MES : Matières En Suspension
MLP : Multi Layer Perceptron
PMC : Perceptron MultiCouche
NTU : Unité néphélométrique de turbidité
NFU : Néphelémétrie (unité de turbidité)
RàPC : Raisonnement à Partir de Cas
ML : Machine Learning
RDF : Reconnaissance Des Formes
EEG : électroencéphalogramme
ECG : électrocardiogramme
TAL : Traitement Automatique des Langues
RST : Rotation, Scale, Translation
SVD : Décomposition en Valeur Singulière
ACI : Analyse en Composantes Indépendantes
LDA : Analyse Discriminante Linéaire
EQMA : L'erreur d'entraînement (Erreur Quadratique Moyenne)
ROM : Random Optimisation- Method
ACP : Analyse en Composantes Principales
FCM : Méthode de la logique floue
PH : potentiel Hydrogène
T° : Température
C : Conductivité
TU : Turbidité
OD : Oxygène Dissous
UV : Ultra Violets
C : concentration (mg)
T : un temps de contact (Min/L)
Al : Aluminium
Sb : Antimoine
As : Arsenic
Ba : Baryum
Be : Béryllium
B : Bore
Cd : Cadmium
Cl : Chlore
Cr⁺³ : Chrome
Cu²⁺ : Cuivre
F⁻ : Fluorure
CaCO₃ : carbonat de calcium

Fe : Fer
Pb : Plomb
Mn : Manganèse
Hg : Mercure
Mb : Molybdène
Ni : Nickel
NO₃ : Nitrate
Se : Sélénium
Ag : Argent
Na : Sodium
U : Uranium
Zn : Zinc
CO₂ : dioxyde de carbone
CaCO₃ : Carbonate de calcium
CaO₂ : Peroxyde de calcium
HCO₃⁻ : Bicarbonates
H⁺ : ions hydrogènes
MO : Matières organiques
T : Ensemble d'échantillons d'entraînement
x_i : Vecteur caractéristique de dimension d
y_i : Étiquette de classe
N : Nombre d'échantillons
W : vecteur de pondération
w_l : Poids associé à la caractéristique l
κ(z) : Fonction du noyau
σ : Largeur du noyau
ξ(w) : La vraie précision de leave-one-out
λ : Paramètre de régularisation
α : Longueur l'étape initiale
η : Petite constante positive

y : Sortie du neurone
x_i : Signaux d'entrée
w_i : Poids d'auto-connexion du neurone i de la couche d'entrée.
f(y) : Fonction d'activation
ei : Cellules d'entrées
yj : Neurones d'activation de la couche cachée
zk : Neurones d'activation de la couche sortie (sortie réelle)
vji : Les poids entre la couche d'entrée et la couche cachée
wkj : Les poids entre la couche cachée et la couche de sortie
x₀ et y₀ : Le biais de réseaux de neurones
sk : La sortie désirée
E_k : Rétro-propagation de l'erreur
F_j : L'erreur de chaque neurone de la couche cachée

INTRODUCTION GENERALE

Bien que souvent perçue comme assez ordinaire, l'eau est la substance la plus remarquable et est la principale composante de notre corps, elle est à l'origine de la vie. Les planètes et les êtres vivants sont essentiellement constitués d'eau. La quantité d'eau à la surface de la terre est constante, cependant 97.23% de celle-ci est en réalité « salée et impropre à une consommation directe. Il faut insister sur le fait que le problème de l'eau n'est pas un problème de quantité, mais un problème de flux. Aujourd'hui dans le monde, 1.2 milliard de personnes n'ont pas accès à une source d'eau potable, et 2.9 milliards ne disposent pas d'installations sanitaires convenables. En 1998, les maladies d'origine hydrique ont tué quelques 3.4 millions de personnes, surtout des enfants. Parmi ces maladies, les plus meurtrières sont la Diarrhée (2.21 millions de victimes), le Malaria (1.11 million), la trypanosomiase, les infections par vers intestinaux, la dengue et la bilharziose [1].

En raison de l'importance évidente de l'eau, il est nécessaire de réfléchir à des solutions pour la contrôler et la surveiller dont le but d'avoir une eau avec une plus grande qualité et avec un coût de production plus faible. Ces considérations motivent les efforts importants dans le développement de nouvelles méthodes et techniques de contrôle et de surveillance [2].

La surveillance de potabilité de l'eau peut être mise en pratique selon deux types de méthodes : classique et moderne. Les méthodes classiques sont déterminées par une analyse chimique effectuée au laboratoire, cette méthode nécessite un temps d'analyse relativement important et peut donc être difficilement intégrée dans un système de surveillance et de diagnostic en temps réel de l'unité. En outre les méthodes modernes qui ont l'avantage de pouvoir effectuer un contrôle automatique permanent en temps réel, sont mieux placées pour être une alternative plus efficace. Une surveillance permanente des divers procédés de traitement et des paramètres relatifs à la qualité de l'eau est devenue nécessaire, ou des systèmes de contrôle automatique infaillibles sont impératifs [3].

Durant ces dernières années, d'importants efforts ont été déployés dans le développement de méthodes de contrôle et de surveillance automatique des eaux potables ; ces méthodes dite de haut niveau, disposent d'outils qui sont plutôt orientés vers la communication avec un opérateur expert [3]. Celle-ci représente les techniques de l'intelligence artificielle (IA) qui servent comme outil de base pour l'aide à la décision. Leur réponse est plus élaborée et peut être obtenue soit à partir de données brutes venant directement des variables de surveillance, ou à partir de données traitées venant des sorties de traitements de bas niveau [4].

Il est logique de supposer que le problème de contrôle et de surveillance de l'eau brute peut être vu comme un problème de reconnaissance de formes, où les formes représentent l'ensemble des observations ou mesures liées aux caractéristiques de celle-ci. Parmi les techniques d'intelligence artificielle (IA) connues dans ce domaine, on trouve « les Réseaux des Neurones Artificiels : RNA ». Cette technique se démarque des autres outils par leur capacité d'apprentissage et de généralisation. Les différents paramètres physico-chimiques exploités dans le traitement de l'eau, tels que le pH, la température (T°), la conductivité (C), l'oxygène dissous et les matières en suspension (MES),...etc, sont transformés en signaux électrique à partir d'une fusion de données multi- sensorielle et transmis vers une station de contrôle qui assure l'acquisition et le traitement des données. La technique devant être utilisée au niveau du système de décision, doit pouvoir effectuer la classification et la séparation de ces données en trois classes (qualité de l'eau très bonne, qualité de l'eau moyenne, qualité de l'eau médiocre), des travaux ont effectuées la classification binaire (deux classes) dans ce domaine avec succès. Dans ce contexte, nous allons choisir une méthode de réduction d'une analyse statistique Analyse en Composantes de voisinage en anglais (Neighborhood Components Analysis : NCA), permettant de déterminer les corrélations existantes entre les variables caractéristiques de l'eau brute puis de ne conserver que les caractéristiques apportant réellement une information pertinente. C'est qu'on peut utiliser comme des variables d'entrée réduit dans un autre système de surveillance.

Le travail présenté dans le cadre de ce mémoire a pour objectif d'étude de la mise en œuvre de deux techniques, le premier est une technique de sélection des paramètres s'appelle NCA. Elle permet de voir l'intérêt de la réduction de dimension pour

minimiser le coût des capteurs des paramètres de l'eau. La deuxième est une technique d'apprentissage statistique (RNA) appliquée dans le domaine du contrôle et de surveillance des eaux potables. Une étude en simulation sera effectuée pour valider, évaluer et comparer les performances de ces modèles dans le but d'avoir un choix décisif adapté au problème de contrôle et de surveillance de la qualité des eaux potables.

Le travail réalisé est axé autour de trois chapitres qui sont présentés comme suit : Le premier chapitre présente une généralité sur les eaux potables. Dans cette section, nous parlerons tout spécialement du processus de traitement des eaux potables, en l'introduisant par des généralités sur leur traitement, leur cycle, les ressources naturelles, ainsi que les normes à appliquer. Nous terminerons ce chapitre par les méthodes de surveillance de l'eau potable.

Dans le deuxième chapitre qui se divise en trois parties: Nous allons aborder à sa première partie une présentation générale sur les systèmes de reconnaissance des formes et son schéma général. Dans la deuxième partie, nous allons aborder la réduction de la dimensionnalité des données d'une façon générale, son objectif ainsi que les différents types. Dans ce contexte, nous allons choisir l'une des méthodes de réduction «NCA». La dernière partie sera dédiée à une présentation sur les classifieur et en particulier les réseaux de neurones artificiels de type MLP (Multi Layer Perceptron). On commencera par une description sur le neurone formel, puis les architectures et l'apprentissage des réseaux. Nous terminerons ce chapitre par la conception d'un réseau de neurones.

Enfin, le troisième est le dernier chapitre va présenter la base des données et les différents paramètres caractéristiques de l'eau. Ainsi, nous allons décrire la mise en œuvre de la technique RNA (avec et sans NCA) appliquée au contrôle de potabilité de l'eau. Une discussion sur les résultats de simulation obtenus sera donnée. Nous concluons cette étude par l'évaluation des performances de l'approche proposée.

Finalement, une conclusion générale retrace les différentes étapes réalisées dans ce mémoire et souligne les perspectives envisagées aux futurs étudiants.

Chapitre I

Généralités sur les eaux propres

I.1. INTRODUCTION

Le but principal d'une station de traitement d'eau potable est de fournir un produit qui satisfait à un ensemble des normes de qualité à un prix raisonnable pour le consommateur. A cet effet, le traitement de l'eau brute est un passage obligé dans le processus de production de l'eau potable. Le but principal dans les plupart des usines de production, est de connaître les différents paramètres descripteurs de la qualité de l'eau brute pour décider de son état, et chercher par la suite les méthodes pour la rendre potable. La surveillance permanente de la qualité de cette ressource à travers des mesures qualifiables et quantifiables, ainsi que du fonctionnement des installations de traitement, est donc exigée afin d'éviter de ralentir la production et assurer un niveau de qualité très élevé et uniforme.

Ce premier chapitre est consacré au processus de contrôle de traitement de l'eau brute. Il s'agit de donner quelques généralités à propos de cette ressource naturelle qu'est l'eau, ainsi que les outils et moyens mis en œuvre pour son traitement. La chaîne de traitement et de production d'eau potable la plus courante est de ce fait présentée. Les différentes méthodes de surveillance sont citées. Notre application est enfin présentée en fin de chapitre, et le principe de l'approche adoptée dans le procédé de surveillance proposée.

I.2. GÉNÉRALITÉS DES EAUX

I.2.1. Cycle de l'eau

L'eau est très présente sur notre planète, les océans recouvrant près des trois quarts de la surface terrestre (70%). La totalité de l'eau sur terre représente un volume de 1,4 milliard de km³, sous forme liquide, solide ou gazeuse [5]. Entre les grands réservoirs d'eau de l'hydrosphère qui sont les mers et océans, les eaux continentales (superficielles et souterraines), l'atmosphère et la biosphère, l'échange d'eau est permanent et forme ce qu'on appelle le cycle externe de l'eau. Le moteur de ce cycle est l'énergie solaire qu'il rayonne, il active et maintient constamment les masses d'eau en mouvement (condensation, évaporation,...). Chauffée par le soleil, l'eau des océans, des mers, des rivières et des lacs s'évapore. A ce volume s'ajoute la vapeur d'eau issue de l'évapotranspiration des plantes. Toute cette vapeur d'eau monte dans l'atmosphère et se refroidit. Elle se condense alors et forme des gouttes ou des cristaux. Les nuages ainsi

constitués se déplacent avec le vent. Ils laissent échapper leur contenu sous forme de précipitations (pluie, neige, grêle) ou de brouillard. La plus grande partie de l'eau tombe directement dans les océans. Le reste atteint les continents, où 60% s'évapore à nouveau, seul 25% s'infiltrate dans la terre et alimente les nappes souterraines, en fin 15% ruisselle et rejoint les cours d'eau, pour retrouver en suite l'océan, et le cycle recommence. L'évaporation, l'action du vent, la condensation, les précipitations, le ruissellement et l'infiltration, sont les 6 phases du cycle de l'eau (figure I.1.). Ce dernier se divise en deux parties intimement liées :

- Une partie atmosphérique qui concerne la circulation de l'eau dans l'atmosphère, sous forme de vapeur d'eau essentiellement.
- Une partie terrestre qui concerne l'écoulement de l'eau sur les continents, qu'il soit superficiel ou souterrain [1].

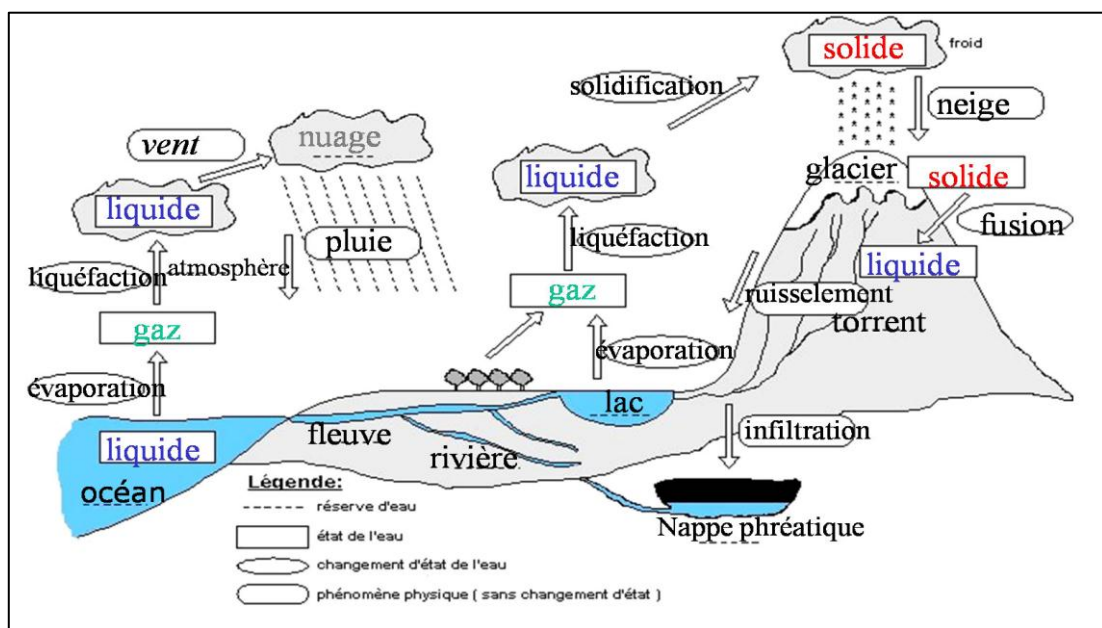


Fig. I.1. Cycle des eaux

I.2.2. Ressources des eaux

Il y a un total de trois sources naturelles d'eau. Ils sont catégorisés comme: l'eau de pluie, l'eau souterraine et l'eau de surface. Voici un regard sur chacun et ce qui fournit ces sources :

I.2.2.1. Les eaux de pluie

L'eau de pluie comprend d'autres sources telles que la neige et d'autres types de précipitations. Les précipitations qui tombent sur la Terre constituent une partie importante du cycle de l'eau, car elles reconstituent à la fois les sources d'eau de surface et souterraines [6]. Ce type de source peuvent être collecté à partir des toitures des maisons dans des récipients ou dans des impluviums. A l'origine ces eaux sont pures sur le plan microbiologique, mais sur le plan chimique, il leur manque souvent certains éléments indispensables à la santé comme le sodium, magnésium, manganèse et le fer [7].

I.2.2.2. Les eaux de surface

Composées d'eaux de mer, de fleuve, de rivière, de marigot, ces eaux couvrent la terre. La terre « planète bleue » en raison de la présence d'eau, 97,5% de celle-ci consiste toutefois en eau salée dont l'essentiel est dans les océans et 2,5% seulement en eau douce (journée mondiale de l'eau 2003). Grossies par les eaux de ruissellement elles reçoivent toutes sortes de déchets contenant des germes nuisibles pour la santé [7].

Estimation de la distribution globale de l'eau :				
Source d'eau	Volume d'eau (km ³)	Volume d'eau (milles ³)	% d'eau douce	% d'eau totale
Océans, mers & baies	1,338,000,000	321,000,000	--	96.5 %
Calottes glaciaires, glaciers et neiges éternelles	24,064,000	5,773,000	68.7 %	1.74 %
Eau souterraine	23,400,000	5,614,000	--	1.7 %
douce	10,530,000	2,526,000	30.1 %	0.76 %
saline	12,870,000	3,088,000	--	0.94 %
Humidité du sol	16,500	3,959	0.05 %	0.001 %
Hydrolaccolithe & pergélisol	300,000	71,970	0.86 %	0.022 %
Lacs	176,400	42,320	--	0.013 %
d'eau douce	91,000	21,830	0.26 %	0.007 %
d'eau saline	85,400	20,490	--	0.006 %
Atmosphère	12,900	3,095	0.04 %	0.001 %
Eau marécageuse	11,470	2,752	0.03 %	0.0008 %
Rivières	2,120	509	0.006 %	0.0002 %
Eau biologique	1,120	269	0.003 %	0.0001 %
Total	1,386,000,000	332,500,000	-	100 %

Source: Gleick, P. H., 1996: Water resources. In Encyclopedia of Climate and Weather, ed. by S. H. Schneider, Oxford University Press, New York, vol. 2, pp.817-823.

Tab. I.1. Estimation de la distribution globale de l'eau.

I.2.2.3. Les eaux souterraines

Ce sont les sources d'eau que vous ne pouvez pas voir ou ne pouvez pas accéder facilement. Le rôle important que jouent les sources d'eau souterraines dans le cycle de l'eau est qu'elles sont les principales sources de consommation humaine. Ces plans d'eau souterrains sont souvent accessibles par des puits et lorsque la pression sous la surface est trop importante, cette eau est poussée vers le haut et vers l'extérieur à travers les sources [6].

I.2.3. Importance de l'eau

L'eau est le principal constituant des êtres vivants et l'élément indispensable à toute forme de vie. Sans eau, aucun organisme, qu'il soit végétal ou animal, simple ou complexe, petit ou gros, ne peut vivre[2].

L'eau était très importante pour les êtres-humains pendant des milliers d'années. Sans eau, il n'y aurait pas de vie sur terre. Nous utilisons de l'eau dans nos maisons pour cuisiner, nous laver et laver la vaisselle [8].

L'eau est utilisée pour cultiver de la nourriture. Les végétaux sont essentiellement formés d'eau. Celle-ci sert à transporter les matières élaborées par la plante et les éléments minéraux nécessaires à sa nourriture. La plante puise dans le sol l'eau et les sels minéraux qui lui permettent d'assurer son développement normal (nutrition, croissance et reproduction) [8].

L'eau est le principal constituant du corps humain. Elle est indispensable pour assurer les différentes fonctions métaboliques. Elle maintient le corps humain et le protéger contre les maladies [8].

Les industries et les usines utilisent également de l'eau. Les fruits et légumes doivent être nettoyés avant de pouvoir être transformés et vendus dans les supermarchés. L'eau est utilisée pour le refroidissement dans de nombreux domaines, par exemple dans la production d'acier [8].

De nombreux pays à travers le monde utilisent l'eau pour produire de l'énergie. Les centrales électriques brûlent du charbon qui transforme l'eau en vapeur. Les pays avec de nombreuses montagnes et rivières utilisent le pouvoir de l'eau pour produire de l'électricité. L'eau est importante pour notre temps libre. Les gens s'amuse dans les

stations balnéaires ou lors de croisières. Le transport a d'abord été effectué sur les voies navigables [8].

Les civilisations anciennes échangeaient des biens à travers la Méditerranée. Aujourd'hui, le pétrole, le charbon, le blé et d'autres produits sont transportés sur les voies navigables [8].

I.2.4. Le problème d'eaux

Durant les dernières décennies, la qualité des eaux de surface s'est fortement dégradée en raison des activités anthropiques telles que l'agriculture, le changement des usages du sol, l'imperméabilisation des surfaces à l'origine d'un ruissellement plus important et plus chargé en polluants, les rejets d'eaux usées non traitées, et pour une part difficile à évaluer, le changement climatique qui influence la température et les caractéristiques chimiques des plans d'eau.

Aujourd'hui Plus de 80% des égouts des pays en voie de développement se déversent sans traitement, polluant les rivières, les lacs et les zones côtières. Bien que selon les estimations les populations rurales d'Asie devraient rester stables au cours des 20 prochaines années, les populations urbaines vont certainement augmenter de 60% avant 2025, ce qui entraîne des perspectives de pénurie d'eau potable.

Globalement, le problème de qualité de l'eau le plus préoccupant est l'eutrophisation, résultat d'une forte concentration en nutriments (principalement phosphore et nitrogène), ce qui réduit significativement l'exploitation utile de l'eau.

Suivi de l'impact des changements globaux, en particulier du changement climatique, entraînent une forte demande des systèmes performants de suivi de la qualité de l'eau [23].

I.2.5. L'eau potable

Une eau potable est une eau qui doit satisfaire un certain nombre de caractéristiques la rendant propre à la consommation humaine, elle ne doit pas renfermer de substances chimiques en quantité ni de germes nocifs pour la santé. En réalité, elle est chargée d'une très grande variété de matières dissoutes inertes ou vivantes, des substances minérales ou organiques, des micro-organismes ainsi que des particules en suspension et gaz dissous. Les standards servant de référence dans ce domaine sont différents et varient selon l'autorité en charge de leur définition. Le concept de

potabilité est donc une notion propre à chaque pays, résultat du contexte culturel et historique local, il détermine la question de l'accès à l'eau, puisque une eau de bonne qualité est essentielle au développement économique et humain [2].

I.2.6. Normes de la qualité des eaux

Une eau de consommation ne doit pas contenir de germes des maladies à transport hydrique, de substances toxiques ni de quantité excessive de matières minérales et organiques. Elle doit par ailleurs, être limpide, incolore et ne posséder aucun goût ou odeur désagréable. En outre l'eau potable doit contenir sans excès un certain nombre d'éléments minéraux dont la présence lui confère une saveur agréable à l'exclusion de ceux qui seraient l'indice d'une contamination ainsi que toute substance toxique [7].

Elément/ substance	Symbole/ formule	Concentration normalement trouvée dans l'eau de surface	Lignes directrices fixées par l'OMS
Aluminium	Al		0,2 mg/l
Antimoine	Sb	< 4 µg/l	0.02 mg/l
Arsenic	As		0,01 mg/l
Baryum	Ba		0,7 mg/l
Béryllium	Be	< 1 µg/l	Pas de valeur guide
B	B	< 1 mg/l	0.5mg/l
Cadmium	Cd	< 1 µg/l	0,003 mg/l
Chlore	Cl		Pas de valeur mais on peut noter un goût à partir de 250 mg/l
Chrome	Cr ⁺³ , Cr ⁺⁶	< 2 µg/l	chrome total : 0,05 mg/l
Cuivre	Cu ²⁺		2 mg/l
oxygène dissous	O ₂		Pas de valeur guide
Fluorure	F ⁻	< 1,5 mg/l (up to 10)	1,5 mg/l
Dureté	CaCO ₃		200 ppm

Fer	Fe	0,5 - 50 mg/l	Pas de valeur guide
Plomb	Pb		0,01 mg/l
Manganèse	Mn		0,4 mg/l
Mercurure	Hg	< 0,5 µg/l	inorganique : 0,006 mg/l
Molybdène	Mb	< 0,01 mg/l	0,07 mg/l
Nickel	Ni	< 0,02 mg/l	0,07 mg/l
Nitrate et nitrite	NO ₃ , NO ₂		50 et 3 mg/l (exposition à court terme) 0.2 mg/l (exposition à long terme)
Turbidité			Non mentionnée
pH			Pas de valeur guide mais un optimum entre 6.5 et 9.5
Sélénium	Se	<< 0,01 mg/l	0,01 mg/l
Argent	Ag	5 – 50 µg/l	Pas de valeur guide
Sodium	Na	< 20 mg/l	Pas de valeur guide
Sulfate	SO ₄		500 mg/l
Étain inorganique	Sn		Pas de valeur guide : peu toxique
Uranium	U		0.015 mg/l
Zinc	Zn		3 mg/l

Tab. I.2. Normes d'eau potable selon l'OMS (2006).

I.2.7. Importance de l'analyse et du traitement

Une analyse régulière de l'eau est importante pour les raisons suivantes :

- Elle permet de définir les problèmes existants.
- Elle garantit une eau qui convient à l'utilisation prévue.
- Elle garantit une eau potable sûre.
- Elle permet de vérifier l'efficacité du système de traitement.

La qualité d'une réserve d'eau peut changer au fil du temps et même subitement. Si l'apparence, l'odeur et le goût de l'eau restent les mêmes, le changement de qualité risque de passer inaperçu. La seule façon de connaître la salubrité de l'eau potable. Est de la faire analyser. Comme les bactéries, les parasites et les virus nuisibles sont invisibles à l'œil nu, une eau au goût et à l'apparence agréables n'est pas forcément potable. Ces microbes, qui vivent parfois dans l'eau souterraine et de surface, risquent de causer rapidement des maladies chez les humains qui consomment l'eau sans traiter adéquatement. Certains contaminants chimiques que l'on retrouve dans les réserves d'eau peuvent causer des problèmes de santé à long terme, qui n'apparaissent le niveau de salubrité de l'eau et de vérifier si le système de traitement a atteint un degré de purification satisfaisant. Plusieurs analyses disponibles sont utiles pour déterminer la salubrité et la sûreté des réserves d'eau. L'analyse de base de l'eau potable comprend plusieurs aspects d'analyse tels que celui des bactéries coliformes, des nitrates, du pH, du sodium, du chlorure, du fluorure, des sulfates, du fer, du manganèse, des matières totales dissoutes et celui de la dureté [4].

I.3. CHAÎNE ELEMENTAIRE DE LA PRODUCTION DE L'EAU POTABLE

L'industrie de l'eau a une pression croissante pour produire une eau traitée de Plus grande qualité à un coût plus faible. Les eaux à visée de potabilisation pour la consommation humaine sont de différentes natures. Les eaux souterraines constituent 22% des réserves d'eau douce soit environ 1000 milliards de m³ [9]. Elles sont généralement d'excellente qualité physico-chimique et bactériologique. Néanmoins, les terrains traversés en influent fortement la minéralisation. Les eaux de surface se répartissent en eaux courantes ou stockées (stagnantes). Elles sont généralement riches en gaz dissous, en matières en suspension et organiques, ainsi qu'en plancton. Elles sont très sensibles à la pollution minérale et organique de type nitrate et pesticide d'origine agricole.

Dans cette section, nous parlerons plus amplement de traitement des eaux de surface, mais il est certain que certaines eaux souterraines doivent également être traitées. Suivant les circonstances, ces deux types de traitement sont semblables ou différents, mais de toute façon ils présentent des points communs.

Le principal objectif d'une station de production d'eau potable est de fournir un produit qui satisfait à un ensemble de normes de qualité à un prix raisonnable pour le

consommateur. L'efficacité du traitement adopté dépendra de la façon dont sera conduite l'exploitation de l'usine de traitement. Pour atteindre l'objectif souhaité, l'exploitant devra, d'une part, respecter certains principes élémentaires pour assurer le contrôle du processus de traitement et le contrôle de l'eau traitée. D'autre part, il devra disposer d'un certain nombre de moyens techniques et humains [10].

Nous allons présenter, les caractéristiques générales d'une usine de production d'eau potable, la plus complète et la plus courante, tout en détaillant plus spécifiquement le procédé sur lequel porte notre étude. La figure I.2 représente une filière typique de potabilisation appliquée à une eau de surface. Elle comporte des traitements à large spectre d'action tels que prétraitement, oxydation, clarification, désinfection et affinage. Les étapes de déferrisation, démagnétisation, dénitratation sont les principaux traitements spécifiques de l'eau souterraine.

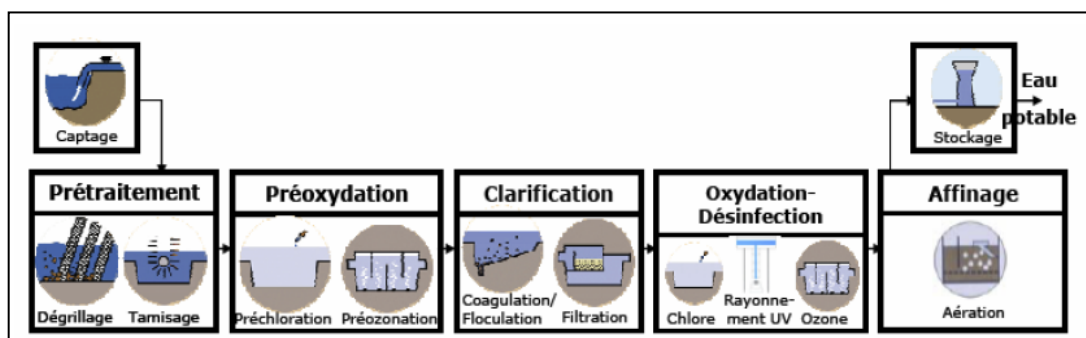


Fig. I.2. Station de production d'eau potable

I.3.1. Prétraitement

Une eau, avant d'être traitée, doit être débarrassée de la plus grande quantité possible d'éléments dont la nature et la dimension constitueraient une gêne pour les traitements ultérieurs. Pour cela, on effectue des prétraitements de l'eau de surface. Dans le cas d'une eau potable, les prétraitements sont principalement de deux types :

- Le dégrillage
- Le tamisage

Le dégrillage, premier poste de traitement, permet de protéger les ouvrages avals de l'arrivée de gros objets susceptibles de provoquer des bouchages dans les différentes unités de traitement. Ceci permet également de séparer et d'évacuer facilement les matières volumineuses charriées par l'eau brute, qui pourraient nuire à l'efficacité des

traitements suivants, ou en compliquer l'exécution. Le dégrillage est avant tout destiné à l'élimination de gros objets : morceaux de bois, ... etc. Le tamisage, quant à lui, permet d'éliminer des objets plus fins que ceux éliminés par le dégrillage. Il s'agit de feuilles ou de morceaux de plastique par exemple [11].

I.3.2. Pré-oxydation

L'oxydation est une opération essentielle à tout traitement des eaux. Elle est toujours incluse en fin de filière au niveau de la désinfection.

A l'issue du prétraitement, on a une eau relativement propre mais qui contient encore des particules colloïdales en suspension. Celles-ci n'ont en elles-mêmes rien de dangereux. Il nous arrive souvent de consommer de l'eau en contenant : le thé, le café, le vin ou le lait qui sont chargés en matières organiques, mais qui s'oxydent spontanément en présence d'air. On va les détruire dans la mesure du possible par une oxydation. Celle-ci peut être faite de trois façons différentes :

- ajoute de Chlore (pré-chloration)
- ajoute de dioxyde de chlore
- ajoute d'ozone (pré-ozonation)

La pré-chloration est effectuée avant le procédé de clarification. Le chlore est le plus réactif et le plus économique, mais il a comme inconvénient de former avec certains micropolluants des composés organochlorés du type chloroforme ou des composés complexes avec les phénols du type chlorophénol dont le goût et l'odeur sont désagréables [12].

On préfère utiliser le dioxyde de chlore qui coûte plus cher mais qui n'a pas les inconvénients de l'oxydation par le chlore cités ci-dessus. Ce type de traitement est cependant réservé à des cas spécifiques. En effet, l'utilisation du dioxyde de chlore présente, lui aussi, des inconvénients non négligeables comme sa décomposition à la lumière, ce qui entraîne une augmentation du taux de traitement à appliquer en période d'ensoleillement. En conclusion, le dioxyde de chlore est un oxydant plus puissant que le chlore qui représente une alternative intéressante à l'utilisation du chlore lorsque celui-ci entraîne des problèmes de qualité d'eau.

Enfin, depuis quinze à vingt ans, on utilise comme oxydant l'ozone, qui non seulement à l'avantage de détruire les matières organiques en cassant les chaînes

moléculaires existantes, mais également à une propriété virulicide très intéressante, propriété que n'a pas le chlore. Généralement utilisée en désinfection finale, cette technique peut être mise en œuvre en oxydation. Elle peut aussi être employée pour l'amélioration de la clarification. L'un des avantages d'une pré-ozonation est l'oxydation des matières organiques, et une élimination plus importante de la couleur. Un autre avantage est la diminution du taux de traitement (taux de coagulant) dans le procédé de clarification. En somme, la pré-ozonation est une solution de substitution à la pré-chloration. On évite ainsi les problèmes liés aux sous-produits de la chloration. Néanmoins, ce procédé ne résout pas tous les problèmes car certaines algues résistent à l'ozone. De plus, son coût reste beaucoup plus élevé que celui au chlore [11].

I.3.3. Clarification

La clarification est l'ensemble des opérations permettant d'éliminer les matières en suspension MES (minérales et organiques) d'une eau brute ainsi que des matières organiques dissoutes [13]. Suivant les concentrations de l'un et de l'autre des différents polluants, on peut être amené à pratiquer des opérations de plus en plus complexes qui vont de la simple filtration avec ou sans réactif jusqu'à la coagulation – floculation – décantation ou flottation – filtration.

La clarification comprend les opérations suivantes :

- ✓ Coagulation
- ✓ Floculation
- ✓ Filtration

La coagulation est l'une des étapes les plus importantes dans le traitement des eaux de surface. 90% des usines de production d'eau potable sont concernées. La difficulté principale est de déterminer la quantité optimale de réactif à injecter en fonction des caractéristiques de l'eau brute.

Un mauvais contrôle de ce procédé peut entraîner une augmentation importante des coûts de fonctionnement et le non-respect des objectifs de qualité en sortie. Cette opération a également une grande influence sur les opérations de décantation et de filtration ultérieures. En revanche, un contrôle efficace peut réduire les coûts de main d'œuvre et de réactifs et améliorer la conformité de la qualité de l'eau traitée.

En résumé, le contrôle de cette opération est donc essentiel pour trois raisons : la maîtrise de la qualité de l'eau traitée en sortie (diminution de la turbidité), le contrôle du coagulant résiduel en sortie (réglementation de plus en plus stricte de la présence de coagulant résiduel dans l'eau traitée) et la diminution des contraintes et des coûts de fonctionnement (coûts des réactifs et des interventions humaines) [11].

I.3.4. Oxydation-Désinfection

La désinfection est l'étape ultime du traitement de l'eau de consommation avant distribution. Elle permet d'éliminer tous les micro-organismes pathogènes présents dans l'eau [10]. Il peut cependant subsister dans l'eau quelques germes banals, car la désinfection n'est pas une stérilisation.

Le principe de la désinfection est de mettre en contact un désinfectant à une certaine concentration pendant un certain temps avec une eau supposée contaminée.

Cette définition fait apparaître trois notions importantes : les désinfectants, le temps de contact et la concentration résiduelle en désinfectant. Une bonne désinfection via les réactifs oxydants demande la combinaison d'une concentration C avec un temps de contact T ; c'est le facteur $C \cdot T$ (mg. Min/L). Cette valeur varie avec les micro-organismes concernés, le type de désinfectant et la température.

Les quatre principaux désinfectants utilisés en production d'eau potable sont les suivants :

- Le chlore
- Le dioxyde de chlore
- L'ozone
- Le rayonnement UV

La concentration en oxydant est pratiquement le seul paramètre sur lequel l'opérateur peut intervenir. Il faut retenir que l'efficacité de la désinfection dépend, en partie, du suivi de la concentration en oxydant. L'évolution de la concentration en oxydant est liée à la demande en oxydant de l'eau. Cette demande dépend de la qualité de l'eau, du pH , des températures (différentes entre l'été et l'hiver), des matières organiques, et de la concentration en ammoniacque [11].

I.3.5. Affinage

Le traitement final traite de la mise à l'équilibre calco-carbonique. L'eau suit un cycle naturel dans lequel les éléments chimiques qu'elle contient évoluent [13]. L'eau de pluie contient naturellement du dioxyde de carbone (CO_2). Quand celle-ci traverse les couches d'humus, riches en acides, elle peut s'enrichir fortement en CO_2 . Lors de sa pénétration dans un sol calcaire, c'est-à-dire riche en carbonate de calcium ($CaCO_3$), elle se charge en Peroxyde calcium CaO_2^+ et en ions bicarbonates HCO_3^- . En fait, le calcium est dissous par l'eau chargée en CO_2 . On dit qu'elle est entartrant ou incrustante. En revanche, quand l'eau de pluie traverse une roche pauvre en calcium (région granitique), elle reste très chargée en CO_2 dissous. Cette eau est, en générale, acide. On dit qu'elle est agressive.

Il y a typiquement deux problèmes distincts : corriger une eau agressive et corriger une eau incrustante. La correction d'une eau agressive peut s'effectuer de plusieurs façons. Premièrement, on peut éliminer le CO_2 par aération. Du fait de l'élimination du CO_2 , le pH augmente et se rapproche du pH d'équilibre. Deuxièmement, on peut ajouter une base à l'eau. L'ajout de base permet d'augmenter le pH et d'atteindre le pH d'équilibre. La correction d'une eau incrustante peut se faire soit par traitement direct soit en réduisant le potentiel d'entartrage par décarbonatation. Le traitement direct correspond à un ajout d'acide [11].

I.4. LES CAPTEURS

Le contrôle permanent et automatique des processus de traitement implique la mesure d'un certain nombre de paramètres à l'aide de capteurs en continu. Ceux-ci peuvent se classer en deux grandes familles: les paramètres usuels et les paramètres spécifiques de l'eau.

Les paramètres usuels sont principalement les débits, les niveaux de liquides ou de solides, les pressions, les températures. Dans toute installation de traitement d'eau la connaissance du débit est impérative. De plus, le traitement d'une eau conduit à lui ajouter un certain nombre de réactifs. Ceux-ci sont généralement stockés soit sous forme liquide dans des bâches, soit sous forme pulvérulente dans des silos. La mesure des débits et des niveaux de produits dans ces unités de stockages permet de connaître à chaque instant l'autonomie prévisionnelle de l'installation et d'établir un planning de livraison des produits. Ces mesures de niveaux s'appliquent également aux différents

réservoirs de stockage (eau traitée, boue, etc.). Dans ces installations, les mesures de pression sont également nombreuses : au refoulement des pompes, dans les filtres, les dégazeurs, ...etc. La bonne conduite d'une installation de filtration nécessite la connaissance permanente du degré de colmatage des lits filtrants qui est généralement donné par une mesure de pression différentielle, ou simplement de pression, dans le cas d'un filtre ouvert à niveau constant. La mesure de température est essentielle, elle est souvent utile sur les réacteurs biologiques, et quelquefois en coagulation-floculation d'eau froide.

La mesure en continu d'un certain nombre de paramètres spécifiques permet de libérer l'opérateur de l'astreinte d'analyse de routine et d'optimiser les traitements en réduisant les temps de réponse. Elle porte principalement sur la turbidité, la conductivité, le pH et la concentration de certains corps dissous. La turbidité est la réduction de la transparence d'un liquide due à la présence de matières en suspension non dissoutes. Elle permet de connaître le degré de pollution physique des eaux à traiter ainsi que la qualité d'une eau livrée à la consommation humaine. Des corrélations sont souvent établies entre turbidité, matières en suspension et couleur. La mesure de la conductivité ou de son inverse, la résistivité, est impérative pour le contrôle des eaux de très haute pureté. La mesure en continu du pH d'une eau, pratiquée à de nombreux stades du traitement, est en particulier utilisable pour le contrôle de la coagulation-floculation, de la désinfection, etc. La mesure en continu de l'oxygène dissous permet de contrôler le maintien en fonctionnement aérobie dans les installations d'épuration par voie biologique, ainsi que les conditions nécessaires à la vie aquatique dans les eaux de surface. La mesure de l'alcalinité et de la dureté de l'eau permet, par le maintien de ces paramètres en deçà de certaines limites, de rester à l'équilibre calco-carbonique de l'eau et donc d'éviter les phénomènes d'entartrage ou de corrosion des canalisations. Le contrôle habituel de la désinfection se fait par la mesure de la quantité résiduelle de l'agent désinfectant : chlore, ozone, ...etc [10].

I.4.1. Mesure des paramètres physico-chimiques

On peut distinguer deux types de paramètres qui sont :

I.4.1.1. Mesure des paramètres usuels

On peut citer quelques paramètres usuels qui sont essentiellement les débits, les niveaux de liquide ou de solide, la pression, et la température.

I.4.1.2. Mesure des paramètres spécifiques

Dans les appareils utilisés pour la mesure des paramètres spécifiques de l'eau, les différentes méthodes d'analyse sont mises en œuvre de façon automatique, en particulier : la néphélogéométrie (mesure de turbidité), la mesure de résistivité ou de conductivité, la potentiométrie (mesure de pH), l'ampèremètre (mesure de concentration en agent oxydant, chlore, ozone), la photo-colorimétrie et la titrimétrie (mesure de la concentration de certaines substances dissoutes dans l'eau) [3]. On peut classer ces différents appareils en deux grandes catégories : celle des capteurs physiques et celle des analyseurs chimiques qui réalisent, préalablement à toute mesure, une ou plusieurs réactions chimiques.

I.4.2. Capteurs physiques

Mesure de la turbidité : la mesure de turbidité de l'eau correspond à une mesure optique des particules en suspension dans l'eau qui lui donnent un aspect trouble. L'unité employée est appelée unité néphélogéométrique de turbidité (NTU). Les particules sont d'origines diverses : Argiles, limons, organismes microscopiques, dépôts dans les canalisations, corrosion ...etc. Les risques sanitaires peuvent être liés à la présence de ces particules car elles permettent aux bactéries et aux virus de se fixer et d'être ainsi protégés de l'action des désinfectants. L'amélioration peut être obtenue par filtration ou coagulation. La prise d'eau s'effectuant dans une nappe peu profonde, donc sur une eau peu filtrée naturellement, la turbidité de l'eau peut varier selon les pluies (et aussi selon les travaux effectués sur le réseau). Elle est la plupart du temps comprise entre 0.1 et 0.3, mais souvent dépasse ces valeurs limites. En France par exemple elle est égale à 2 ; la valeur maximale admissible européenne est de 4 [11].

Donc, le turbidimètre mesure la quantité de lumière diffusée par un échantillon d'eau brute du fait de la présence de particules dans l'eau. Cette valeur est directement proportionnelle à la turbidité de l'échantillon mesuré. Un faisceau lumineux vient toucher la surface sous une incidence telle que ni lui-même ni le faisceau réfléchi ne peut impressionner une cellule photorésistante placée sensiblement perpendiculairement au faisceau incident. Par contre, la lumière diffusée par les particules en suspension vient modifier d'autant plus l'éclairement de la cellule que leur nombre est élevé, ce qui permet d'obtenir la mesure de la turbidité de cette eau [5].

Mesure de la conductivité : La conductivité mesure la capacité de l'eau à conduire le courant entre deux électrodes. La plupart des matières dissoutes dans l'eau se trouvent sous forme d'ions chargés électriquement. La mesure de la conductivité permet donc d'apprécier la quantité de sels dissous dans l'eau. La conductivité est également fonction de la température de l'eau : elle est plus importante lorsque la température augmente. Les résultats de mesure doivent donc être présentés en termes de conductivité équivalente à 20 ou 25°C. Les appareils de mesure utilisés sur le terrain effectuent en général automatiquement cette conversion. Ce paramètre doit impérativement être mesuré sur le terrain. La procédure est simple et permet d'obtenir une information très utile pour caractériser l'eau. Comme la température, des contrastes de conductivité permettent de mettre en évidence des pollutions, des zones de mélanges ou d'infiltration...etc. La conductivité est également l'un des moyens de valider les analyses physicochimiques de l'eau : la valeur mesurée sur le terrain doit être comparable à celle mesurée au laboratoire [14].

Mesure de pH : Le pH (potentiel Hydrogène) mesure la concentration en ions H⁺ de l'eau. Il traduit ainsi la balance entre acide et base sur une échelle de 0 à 14, 7 étant le pH de neutralité. Ce paramètre caractérise un grand nombre d'équilibre physico-chimique et dépend de facteurs multiples, dont l'origine de l'eau. Le pH doit être impérativement mesuré sur le terrain à l'aide d'un pH-mètre ou par colorimétrie [14].

pH < 5	Acidité forte => présence d'acides minéraux ou organiques dans les eaux Naturelles
pH = 7	pH neutre
5,5 < pH < 8	Majorité des eaux souterraines
7 < pH < 8	Neutralité approchée => majorité des eaux de surface
pH = 8	Alcalinité forte, évaporation intense

Tab. I.3. Classification des eaux d'après leur pH

Mesure d'oxygène dissous : l'ampèremètre est utilisé industriellement en traitement des eaux pour la mesure en continu de la concentration en agents oxydants et met en œuvre une méthode simplifiée d'analyse par ampèremètre. La cellule de mesure, qui est alimentée à débit constant en eau à analyser, comporte une

cathode inattaquable, par exemple en platine, et une anode qui peut-être en cuivre, en cadmium, en argent, etc. en l'absence d'agent oxydant, la pile ainsi formée est polarisée que par un courant très faible. Sa dépolarisation et, par conséquent, l'intensité du courant qu'elle débite sont sensiblement proportionnelles à la concentration de l'agent oxydant qui vient se réduire à la cathode. On mesure ainsi la concentration en chlore, ozone, oxygène d'une eau. L'inconvénient de ces appareils réside dans le fait qu'ils mesurent la somme des agents oxydants et qu'ils ne peuvent être vraiment utilisés que dans le cas où un seul corps se trouve en solution à concentration variable. L'effet d'un autre corps, éventuellement présent à concentration constante, peut être annulé par action sur le zéro de l'appareil [5].

I.5. METHODES DES SURVILLANCE

Quand on parle de surveillance des eaux potables, il s'agit en fait de connaître l'état de l'eau en continu (à chaque instant) à partir des différents paramètres ayant trait à sa qualité. Traditionnellement, on procède généralement à un certain nombre de mesures utiles des différents paramètres de l'eau brute pour le test de sa qualité tels que : le contrôle bactériologique, le contrôle de désinfection, et le contrôle physico-chimique (pH, température, turbidité, conductivité, oxygène dissous,.....), et pour décider après sur son état propre, et par la suite chercher les techniques et méthodes pour la rendre potable.

Les méthodes de surveillance industrielle telle qu'elles sont présentées dans ce paragraphe sont illustrées sur la figure I.3. L'existence d'un modèle formel ou mathématique de l'équipement détermine la méthode de surveillance utilisée. La surveillance avec modèle se compose essentiellement de deux techniques : méthodes de redondance physique et analytique et méthodes d'estimation paramétrique. D'un autre côté, les méthodes qui ne se basent pas sur l'existence du modèle se divisent en deux catégories : méthodes utilisant des outils statistiques et méthodes de reconnaissance des formes. Les outils statistiques établissent des tests sur les signaux d'acquisition. Ces tests ne sont capables d'assurer que la fonction détection de défaillances. Par contre, les techniques de surveillance par reconnaissance des formes sont plus élaborées par rapport aux simples tests statistiques et sont capables de détecter et de diagnostiquer les défaillances [3].

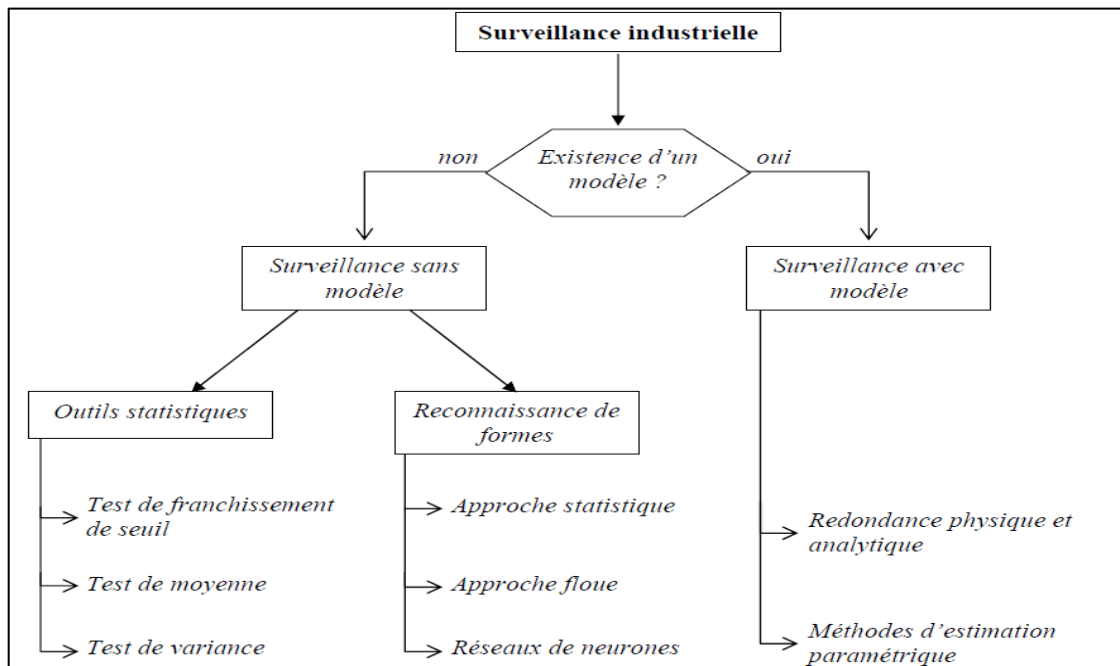


Fig. I.3. Classifications des méthodologies de surveillance industrielle.

I.5.1. Méthodes de surveillance avec modèles

Les méthodes de surveillance avec modèle ont pour principe de comparer les mesures effectuées sur le système aux informations fournies par le modèle [15]. Tout écart est alors synonyme d'une défaillance. Les outils de la théorie de la décision sont ensuite utilisés pour déterminer si cet écart est dû à des aléas normaux comme, par exemple, le bruit de mesure ou s'il traduit une défaillance du système. Ces méthodes peuvent être séparées en deux techniques : techniques de redondance physique et analytique et techniques d'estimation paramétrique. Ces deux techniques seront présentées brièvement. Toutefois, pour plus de détails, nous renvoyons le lecteur aux références suivantes :

I.5.1.1 Redondances physiques et analytiques

a. Redondances physiques

Afin de fiabiliser la détection des défaillances à partir des signaux mesurés, il faut un moyen pour distinguer les défaillances capteurs des défaillances système. La méthode la plus simple consiste à utiliser la redondance physique. Il s'agit de doubler ou tripler des composantes de mesure du système. Si ces composantes identiques placées dans le même environnement émettent des signaux identiques, on considère que

ces composants sont dans un état de fonctionnement nominal et, dans le cas contraire, on considère qu'une défaillance capteur s'est produite dans au moins une des composantes [16]. Cette méthode par redondance physique a l'avantage d'être conceptuellement simple mais est coûteuse à être mise en œuvre et conduit à des installations encombrantes. Elle est, par conséquent, utilisée uniquement pour la surveillance des sous-ensembles critiques d'un système. Un autre inconvénient est que les composantes identiques fabriquées dans la même série peuvent se dégrader de la même façon et tomber en panne en même temps. Pour pallier ce dernier inconvénient, on peut utiliser des composantes différentes qui remplissent la même fonction.

b. Redondances analytiques

Les méthodes de redondance analytique nécessitent un modèle du système à surveiller. Ce modèle comprend un certain nombre de paramètres dont les valeurs sont supposées connues lors du fonctionnement nominal. Dans la mesure où la surveillance est établie à partir des mesures échantillonnées des grandeurs observables du système, la modélisation de ce dernier sous forme discrète semble être raisonnable. Le but des méthodes de redondance analytique est d'estimer l'état du système afin de le comparer à son état réel. L'estimation de l'état du système peut être réalisée soit à l'aide de techniques d'estimation d'état, soit par obtention de relations de redondance analytique. La théorie de la décision est ensuite utilisée pour déterminer si l'écart observé est dû à des aléas normaux du fonctionnement ou à des défaillances [3].

I.5.1.2. Méthodes d'estimation paramétrique

Les méthodes d'estimation paramétrique supposent l'existence d'un modèle paramétrique décrivant le comportement du système et que les valeurs de ces paramètres en fonctionnement nominal soient connues. Elles consistent alors à identifier les paramètres caractérisant le fonctionnement réel, à partir de mesures des entrées et des sorties du système [17].

On dispose ainsi d'une estimation des paramètres du modèle, effectuée à partir des mesures prises sur le système et de leurs valeurs théoriques. Pour détecter l'apparition de défaillances dans le système, il faut effectuer la comparaison entre les paramètres estimés et les paramètres théoriques. Comme pour les méthodes de redondance analytique, la théorie de la décision sert alors à déterminer si l'écart observé est dû à des aléas normaux du fonctionnement ou à des défaillances. La différence entre

les méthodes de redondance analytique et les méthodes d'estimation paramétrique est qu'on effectue, pour les premières, la comparaison entre l'état estimé et l'état théorique du système, alors que pour les secondes, on compare les paramètres estimés aux paramètres théoriques du système.

Les méthodes d'estimation paramétrique requièrent donc l'élaboration d'un modèle dynamique précis du système à surveiller. Ceci restreint leur utilisation à des procédés bien définis. Nous pouvons citer à titre d'exemple les travaux réalisés par [18] qui se basent exactement sur la technique d'estimation paramétrique. Un réseau de neurones sert à estimer les paramètres physiques d'une machine-outil à partir de la mesure du courant, tension et vitesse de rotation de l'axe de la machine-outil. Les valeurs estimées sont utilisées comme base pour la détection et le diagnostic des défaillances.

I.5.2. Méthode de surveillance sans modèle

Pour plusieurs applications industrielles le modèle de procédé est difficile à construire. Une large quantité de données enregistrées sur le fonctionnement du système (normal et au cours de défaillance) est nécessaire. Deux méthodes existent dans ce cas : surveillance avec des outils statistiques de traitement du signal, et surveillance par des méthodes à partir de données historiques.

I.5.2.1. Surveillance utilisant des outils statistiques de traitement du signal

Cette technique de surveillance suppose que les signaux provenant de la chaîne de mesure possèdent certaines propriétés statistiques. Ces propriétés sont identifiées par des tests qui permettent de vérifier si elles sont présentes dans un échantillon des signaux mesurés prélevés en mode de fonctionnement normal. Parmi les plus importants, on trouve : le test de franchissement de seuil et les tests de moyenne et de variance [19, 20].

I.5.2.2. Surveillance par des méthodes à partir de données historiques

Ces méthodes s'appuient largement sur les techniques de l'Intelligence Artificielle (IA) qui permettent de pallier la complexité des systèmes à diagnostiquer. Généralement, l'Intelligence Artificielle est relativement bien adaptée aux problèmes de diagnostic. Une première approche consiste à considérer la surveillance par reconnaissance de formes suivant une approche probabiliste, une approche floue ou une

approche neuronale [21]. Globalement, ces méthodes seront regroupées sous l'expression : "Méthodes à partir de données historiques ou Méthodes d'Intelligence Artificielle". Cette classe comporte trois modèles : les méthodes à base de modèles comportementaux, les méthodes de reconnaissance de formes et les méthodes à base de modèles explicatifs.

I.5.2.3. Les méthodes à base de modèles comportementaux

Ces méthodes consistent à simuler le comportement du système à partir d'une modélisation de son comportement. Le plus souvent il s'agit de modèles "de bon fonctionnement" qui, contrairement aux modèles numériques, ne sont pas basés sur la physique du système mais sur une considération en termes de mode de fonctionnement. Ces méthodes regroupent deux principales familles : les réseaux de Pétri et les automates d'états finis, qui seront caractérisées par les termes de modélisation/simulation [22].

I.5.2.4. Les méthodes de reconnaissance de formes

Ces méthodes regroupent les modèles associatifs et les méthodes de reconnaissance dans le sens où elles sont caractérisées par les termes apprentissage et reconnaissance qui s'appliquent aussi bien aux systèmes de reconnaissance de formes qu'aux systèmes à base de règles tels que les systèmes experts. Dans ces méthodes, on retrouve principalement des outils tels que les méthodes statistiques de reconnaissance de formes, les réseaux de neurones artificielles, logique floue, les réseaux neuro-flous, les systèmes experts et le raisonnement à partir de cas (*RàPC*). Les mots clés qui caractérisent mieux ces approches sont : apprentissage/reconnaissance [22].

I.5.2.5. Les méthodes à base de modèles explicatifs

Ces méthodes sont basées sur la formulation des relations entre les défaillances, leurs causes et leurs effets observables. Ces méthodes consistent à trouver la cause d'une défaillance à partir de ses symptômes observables. Parmi ces méthodes, on trouve les outils tels que : les graphes de fluence, les graphes causaux, les graphes contextuels. Le mot clé associé est l'analyse causale [22].

I.6. SURVEILLANCE PAR RECONNAISSANCE DE FORMES

L'approche de surveillance par reconnaissance de formes (ou par analyse des données), permet d'associer un ensemble de mesures effectuées sur le système à des états de fonctionnement connus. Cette fonction permet d'avoir une relation d'un espace de caractéristiques vers un espace de décision de façon à minimiser le risque de mauvaise classification. Les techniques de reconnaissance de formes sont nombreuses et pour certaines, les résultats sont globalement satisfaisants. Ces méthodes supposent qu'aucun modèle n'est disponible. Cependant, elles effectuent les opérations de perception, de classification et de prédiction. Ces méthodes sont basées sur les théories des probabilités et de l'Intelligence Artificielle. La première peut se montrer insuffisante car elle suppose une connaissance a priori de tous les états de fonctionnement et ne prend pas en compte l'évolution du système. La seconde a l'avantage de se baser plutôt sur une phase dite d'apprentissage. Parmi les techniques proposées, on trouve en particulier les systèmes experts, les outils statistiques de reconnaissance de formes, le (RàPC), les réseaux de neurones artificiels (RNA), le SVM, la logique floue et réseaux neuro-flous. La figure I.4. montre l'architecture générale qu'on peut imaginer d'une application de surveillance par reconnaissance de formes. L'expert humain joue un rôle primordial dans ce type d'applications. Toute la phase d'apprentissage supervisée dépend de son analyse des états du système où chaque état est caractérisé par un ensemble de données (formes d'entrée) recueillies sur le système. Après cette phase d'apprentissage, l'algorithme utilisé associera les classes correspondantes représentant ainsi les sorties du système aux formes d'entrée présentées. Le principal avantage de ces techniques est leur capacité d'apprentissage et de généralisation des données reçues en entrée. Ces techniques s'apprentent parfaitement pour une éventuelle mise au point de systèmes de surveillance capables de s'adapter à d'éventuelles extensions et reconfigurations multiples [1].

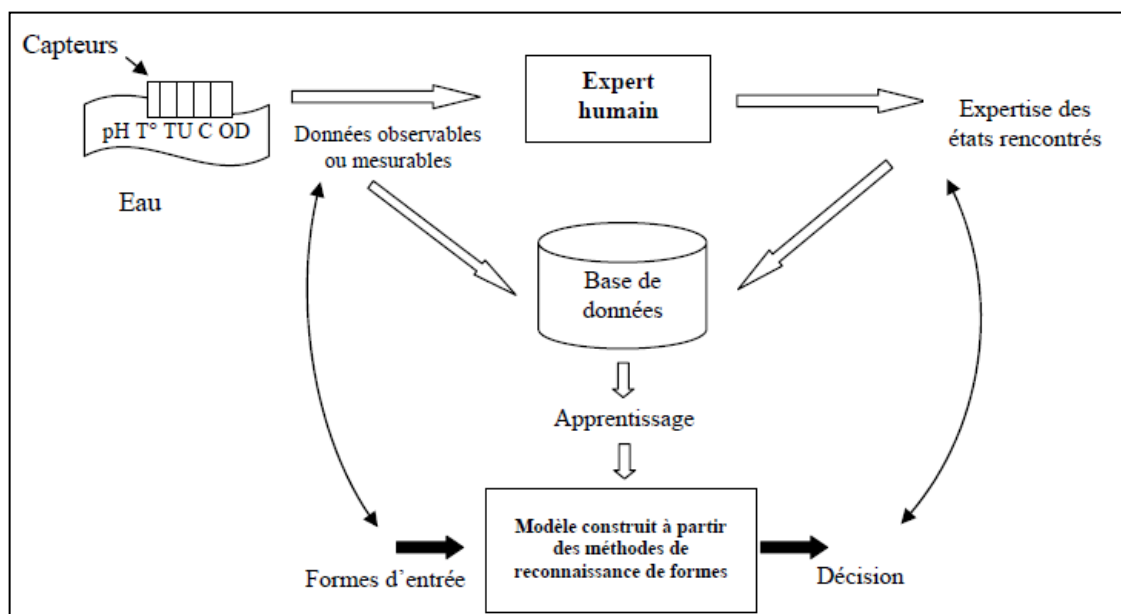


Fig. I.4. Schéma général du système de surveillance par reconnaissance de formes.

I.6.1. Objectif de ce travail

Dans le travail que nous présentons dans ce mémoire, la surveillance de la qualité de l'eau potable peut être vue comme un problème de reconnaissance de formes (apprentissage, classification, décision etc.), où les formes représentent l'ensemble des paramètres relatifs à la qualité de l'eau, et les classes correspondent aux différents états de l'eau.

Cependant, l'importance de ces paramètres est variée d'un paramètre à un autre. Aussi, il arrive parfois qu'un de ces capteurs qui nous fournissent l'information sur la qualité des eaux ne fonctionne pas correctement. Plus de ça, le coût d'installation de plusieurs capteurs est considérable. A ce stade-là, il est très nécessaire de réfléchir à réduire le nombre des paramètres qui s'interviennent à la décision sur la qualité des eaux.

Pour cette raison, l'objet de notre étude est d'exploiter les techniques de la réduction des dimensions dans l'espace caractéristiques pour minimiser le nombre de paramètres à collecter, tout en gardant une performance très élevée au niveau de la décision. La sélection des paramètres nous permet de mettre pour chaque un des eux un poids qui signifie le degré de son pertinence. Finalement, les paramètres les plus informatifs (qui ont les poids les plus hauts) seront considérés dans la phase de la classification.

I.7. CONCLUSION

Ce premier chapitre a servi comme introduction au domaine de contrôle et de surveillance des eaux potables. Les différentes étapes d'une chaîne de traitement sont présentées. Les paramètres ainsi que les capteurs physico-chimiques utilisés comme source d'information ayant trait à la qualité de l'eau ont été particulièrement décrits. De même, les différentes techniques existantes actuellement dans le domaine du contrôle et de surveillance des eaux potables ont été aussi évoquées.

Nous allons présenter dans le chapitre suivant des techniques de reconnaissance de forme utilisées dans le domaine du traitement des eaux propres. Également, et afin de savoir quels sont les données les plus représentatifs dans ce contexte, nous allons décrire une méthode de sélection des paramètres basée sur l'analyse de voisinage.

Chapitre II

Systeme proposé

II.1. INTRODUCTION

Les machines d'apprentissage « Machine Learning : ML » et la classification ont été conçues pour la résolution automatique des problèmes complexes de reconnaissance des formes, où la prise de décisions sur la base d'observations de ces problèmes se fait d'une manière automatique. En effet, le progrès scientifique a permis d'augmenter sensiblement les capacités de mesures et a ainsi augmenté la taille des bases de données. Il est à présent difficile pour un opérateur humain de traiter ces données dans un temps raisonnable. Ce qui a rendu la tâche de prise de décision plus difficile et complexe pour un opérateur humain.

Dans ce chapitre, nous allons aborder les différents modules d'un système de reconnaissance des formes. Ensuite, nous nous intéresserons aux méthodes de réduction de dimension, leurs principes théoriques et leurs applications. Un aperçu de la méthode de « Neighborhood Component Analysis : NCA » sera donné dans ce chapitre. Après cela, nous présentons la classification des données comme une tâche de décision. Pour ce faire, nous allons exploiter la méthode de réseaux de neurones artificiels pour la classification multi-classes.

II.2. SYSTEME DE LA RECONNAISSANCE DES FORMES

II.2.1. Présentation de la reconnaissance des formes

La reconnaissance des formes (ou RDF) est issue de différentes disciplines qui sont les mathématiques (probabilité et statistiques), les sciences de l'ingénieur, l'informatique et l'intelligence artificielle. C'est à partir des années 60 que la reconnaissance des formes est devenue une discipline spécifique. L'extraordinaire développement des ordinateurs ces dernières années a donné un élan à la RDF en permettant des applications temps réel, en particulier dans le domaine des applications visuelles et auditives. Les procédés d'acquisition tels que camera, scanner sont très accessibles, ainsi que des ordinateurs à la fois puissants et bons marchés. Ils permettent le traitement de nombreuses données en un temps raisonnable comme cela est souvent nécessaire en RDF.

L'objectif de la RDF est de réaliser des systèmes informatisés qui simulent les activités humaines de perception, de reconnaissance et de compréhension : par exemple

la reconnaissance de l'écriture, de la parole, interprétation de scènes, robotique, reconnaissance de signaux médicaux EEG (électroencéphalogramme), ECG (électrocardiogramme). Cela implique aussi une certaine pluridisciplinarité pour comprendre l'aspect physique des capteurs, les aspects mathématiques de la classification, ceux relatifs à l'informatique.

Les systèmes de reconnaissance des formes intègrent toute la chaîne perception reconnaissance depuis l'acquisition des données brutes jusqu' à la compréhension élaborée de ces données. Ces dernières ont entretemps subi de nombreuses transformations. De ce fait, la RDF fait appel a des disciplines connexes telles que le traitement du signal et de l'image, l'intelligence artificielle ou le traitement automatique des langues (TAL) [24].

II.2.2. Schéma général d'un système de reconnaissance

Une chaîne de traitement dans un système de reconnaissance comprend plusieurs modules (Figure II.1) et plusieurs espaces de travail. L'objectif de la reconnaissance des formes est de définir une suite d'opérations permettant de passer de l'espace des données ou formes, à l'espace des classes ou la catégorie de la forme est estimée. Ces opérations sont en pratique des procédures informatisées.

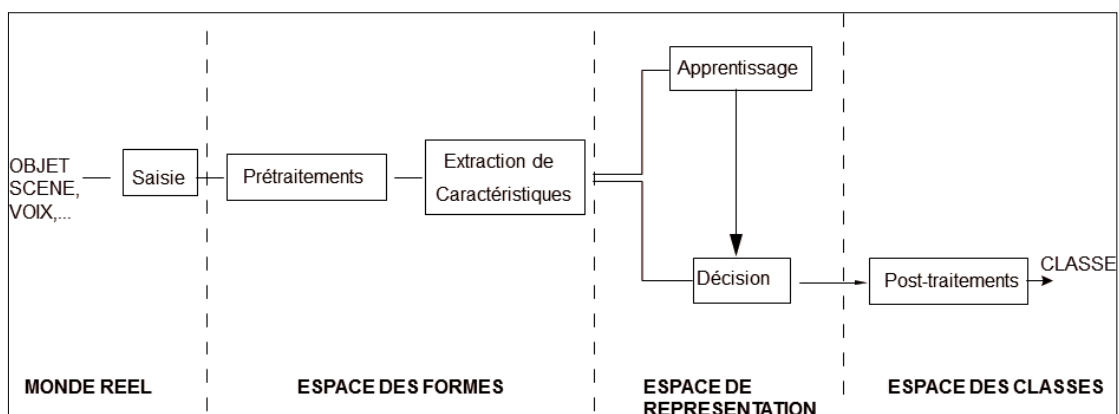


Fig. II.1. Schéma général d'un système de reconnaissance de formes.

Avant d'estimer la classe d'une forme, celle-ci doit être numérisée puis représentée. Les représentations peuvent être diverses: vecteurs ou séquences de vecteurs, arbres ou graphes. La fonction de décision peut être construite à partir de

représentations extraites sur des formes issues d'un ensemble d'apprentissage : c'est la phase dite d'apprentissage. Une décision peut alors être prise pour une forme inconnue, dont on a extrait au préalable sa représentation [24].

II.2.2.1.Système d'acquisition et prétraitement

Suivant la nature du signal, un capteur est nécessaire pour acquérir le signal sous forme numérique ou analogique (il faut alors le convertir en numérique) pour qu'il soit traitable par un système informatisé. On passe ainsi du monde réel au monde des formes ou donnée.

Les prétraitements sont spécifiques à un domaine et ont pour utilité de réduire les bruits de capteurs ou inhérent au signal. Ils peuvent servir aussi de préparation aux phases suivantes en réduisant la variabilité du signal [24].

II.2.2.2.Extraction de caractéristiques

La fonction de décision d'un système de reconnaissance des formes repose largement sur les données. Ces données sont des mesures ou observations, symboliques ou numériques, extraites des formes. Par exemple une température, une longueur, la luminance reçue par un capteur, un voltage, etc. Bien choisir ces caractéristiques et en prendre un nombre suffisant est vital pour développer un bon système de reconnaissance.

Le système de reconnaissance va utiliser ces caractéristiques, en nombre d , pour déterminer la classe d'une forme inconnue. Le système considère K classes différentes. Celles ci sont par exemple des types d'animaux, des lettres, la présence ou absence d'une tumeur, ...etc. Dans la Figure II.2, $K = 3$, $d = 2$ et chaque classe est représentée dans l'espace de représentation (c-à-d l'espace des caractéristiques) par une forme de point différentes. Les peuvent être aussi représentées par des couleurs de points différentes.

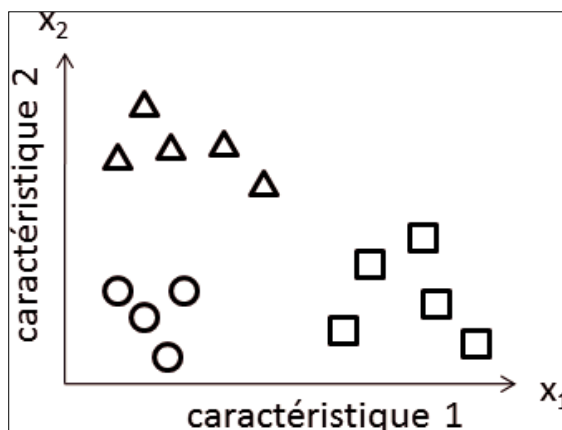


Fig. II.2. Caractéristique extraites sur des formes appartenant a 3classes dans un espace de représentation bidimensionnel $d = 2$ et $K = 3$.

Les caractéristiques peuvent être entachées de bruit ou d'erreurs, elles peuvent être aussi continues ou discrètes. Elles sont choisies pour résoudre un problème précis et conduire à une erreur de classification faible. Ceci sera possible si les caractéristiques sont discriminantes : c-à-d si elles varient peu par miles formes d'une classe donnée et différent autant que possible entre formes de classes distinctes.

Le concepteur cherchera à extraire des caractéristiques de type RST, c-à-d invariantes à la rotation, à l'échelle et à la translation. Il prendra aussi en compte les déformations les plus probables des formes et leurs répercussions dans la chaîne de traitement. L'objectif de la reconnaissance des formes est de regrouper les formes en catégories à partir de leur caractéristiques communes [24].

II.2.2.3. Classification

La classification consiste à calculer une fonction de décision qui transforme les attributs caractérisant les signaux en appartenance à une classe. Ce calcul est effectué en mesurant la ressemblance d'un signal inconnu avec un ensemble de références stockées en mémoire. Ces références sont obtenues lors d'une phase antérieure qualifiée d'apprentissage qui est une étape clé dans le système de reconnaissance. On distingue deux types d'apprentissage : l'apprentissage supervisé et l'apprentissage non supervisé. Dans le cas de l'apprentissage supervisé, chaque signal est doté d'une étiquette indiquant sa classe d'appartenance. La phase d'apprentissage consiste à analyser les ressemblances entre les éléments d'une même classe et les dissemblances entre les

éléments de classes différentes pour en déduire la meilleure partition de l'espace de décision. Dans l'apprentissage non supervisé, on fournit au système un grand nombre de signaux non étiquetés [25]. Le module de classification se charge d'identifier automatiquement les signaux si appartenant à une même classe. Dans notre travail, nous allons utiliser les réseaux de neurones artificiels (RNA) avec un apprentissage supervisé pour la classification de nos données.

II.3. RÉDUCTION DE DIMENSION DES DONNÉES

Le développement technologique et le besoin de superviser des systèmes de plus en plus complexes nécessitent l'analyse de bases de données de taille importante (signaux, images, documents, ...). Toutefois, si dans cette accumulation de données, on est sûr d'avoir une information complète et utile, celle-ci risque d'être "noyée" dans la masse. Ceci pose les problèmes de la structuration des données et de l'extraction des connaissances. En effet, les bases de données sont en général définies par des tableaux à deux dimensions correspondant aux données et aux attributs les caractérisent. Ces deux dimensions peuvent prendre des valeurs très élevées, ce qui peut poser un problème lors du stockage, de l'exploration et de l'analyse. Pour cela, il est important de mettre en place des outils de traitement permettant l'extraction des connaissances sous-jacentes [26]. Les principaux objectifs de la réduction de dimension peuvent être décrits par :

- ✓ identification des attributs pertinents,
- ✓ amélioration de la tâche de classification,
- ✓ facilitation de la visualisation et de la compréhension des données,
- ✓ réduction de l'espace de stockage nécessaire,
- ✓ réduction du temps d'apprentissage et d'utilisation. Toutefois,

L'élimination de certaines informations redondantes ou peu pertinentes peut augmenter l'erreur de classification, vu que ces informations peuvent s'avérer informatives si on les utilise conjointement [27].

L'extraction des connaissances s'effectue selon deux directions, la catégorisation des données (par regroupement en classes) et/ou la réduction de la dimension de l'espace de représentation de ces données (par sélection ou extraction d'attributs). La classification vise à découvrir la structure intrinsèque d'un ensemble de données en formant des groupements qui partagent des caractéristiques similaires [26].

La réduction de la dimensionnalité des données consiste à développer une approche capable de déterminer parmi l'ensemble des variables observées notées N celles réellement nécessaires notées M pour préserver l'information pertinente, la mettre en évidence en la dissociant du bruit avec la possibilité de révéler une structure sous-jacente qui ne sera pas immédiatement apparente dans les données d'origine en haute dimension. La variable N est appelée la dimension d'observation des données et M est la dimension intrinsèque. Notons que la dimension intrinsèque correspond au nombre de variables nécessaires et suffisantes pour représenter efficacement les données. L'exemple classique d'une telle approche est les algorithmes d'analyse en composantes principales et l'analyse en composantes indépendantes [26].

Les deux différences majeures entre la sélection des caractéristiques, et les méthodes du prétraitement non supervisé destinées à projeter les données dans un nouvel espace de dimension inférieur (dans le but de la réduction) sont:

- ✓ Au lieu de choisir un sous-espace de caractéristiques, la projection de données crée de nouvelles dimensions définies par une fonction de toutes les autres caractéristiques.
- ✓ Puis cette même réduction ne considère pas l'étiquette des classes, mais plutôt la donnée point par point [28].

II.3.1. Les méthodes de Réduction de dimension des données

Dans les cas des méthodes non linéaires, nous exploitons les données sans connaissances préalables du modèle. Nous exposerons quelques-unes de ces techniques de prétraitement de données précédant l'analyse postérieure de celles-ci : l'analyse en composantes principales (ACP), la décomposition en valeur singulière (SVD) ainsi que l'analyse en composantes indépendantes (ACI). Contrairement aux méthodes non linéaires, celles-ci imposent des limites à l'analyse des données sous certaines contraintes du modèle. Nous citons à titre d'exemple l'analyse discriminante linéaire (LDA) [28].

Plus récemment, une technique de réduction de dimension, basée sur la probabilité complémentaire du voisinage (les poids), a été proposée comme une méthode efficace pour améliorer la discrimination des données [29]. Cette méthode s'appelle l'analyse en composantes voisines « Neighborhood Component Analysis :

NCA ». Nous allons donc utiliser cette technique pour réduire la dimension de nos données. Une description de son principe sera donnée dans la section suivante.

II.3.1.1. Neighborhood Component Analysis (NCA)

Soit $T = \{(x_1, y_1), \dots, (x_i, y_i), \dots, (x_N, y_N)\}$ l'ensemble d'échantillons d'entraînement, où x_i est un vecteur caractéristique de dimension d , $y_i \in \{1, \dots, C\}$ est son étiquette de classe correspondante et N est un nombre d'échantillons. Le but est de trouver un vecteur de pondération w qui se prête à sélectionner le sous-ensemble d'entités en optimisant la classification des plus proches voisins. En termes de vecteur de pondération w , nous désignons la distance pondérée entre deux échantillons x_i et x_j par :

$$D_w(x_i, x_j) = \sum_{l=1}^d w_l^2 |x_{il} - x_{jl}| \quad (1)$$

Où w_l est un poids associé à la caractéristique l . Pour que la classification des plus proches voisins réussisse, une intuitive et une stratégie efficace est de maximiser la précision leave-one-out classification sur l'ensemble d'apprentissage T . Cependant, en raison de la véritable précision de leave-one-out qui sélectionne le plus proche voisin comme un point de référence de classification, est un non différentiable fonction. Une approximation efficace est que le point de référence est déterminé par une distribution de probabilité. Ici, la probabilité que x_i sélectionne x_j comme point de référence est définie comme :

$$p_{ij} = \begin{cases} \frac{\kappa(D_w(x_i, x_j))}{\sum_{k \neq i} \kappa(D_w(x_i, x_k))} & \text{if } i \neq j \\ 0, & \text{if } i = j \end{cases} \quad (2)$$

Où $\kappa(z) = \exp(-z / \sigma)$ est une fonction du noyau et le La largeur du noyau σ est un paramètre d'entrée qui influence la probabilité de chaque point étant sélectionné comme point de référence. En particulier, si $\sigma \rightarrow 0$, seul le voisin le plus proche de l'échantillon de requête peut être sélectionné comme point de référence. D'autre part, si $\sigma \rightarrow +\infty$, tous les points ont la même chance d'être sélectionné en dehors du point de requête. Sur la base de la définition ci-dessus, la probabilité de la requête x_i pour qu'il soit correctement classifié est donné par:

$$p_i = \sum_j y_{ij} p_{ij} \quad (3)$$

Où $y_{ij} = 1$ si et seulement si $y_i = y_j$ et $y_{ij} = 0$ sinon. Par conséquent, l'approximation de la classification peut s'écrire comme suit:

$$\xi(w) = \frac{1}{N} \sum_i p_i = \frac{1}{N} \sum_i \sum_j y_{ij} p_{ij} \quad (4)$$

Notez que quand $\sigma \rightarrow 0$, $\xi(w)$ est la vraie précision de leave-one-out classification. De plus, pour effectuer la sélection de caractéristiques et atténuer le surajustement, nous introduisons un terme de régularisation et donc obtenir la fonction objective suivante:

$$\xi(w) = \sum_i \sum_j y_{ij} p_{ij} - \lambda \sum_{l=1}^d w_l^2 \quad (5)$$

où $\lambda > 0$ est un paramètre de régularisation qui peut être réglé par validation croisée. Notez que le coefficient $1/N$ dans (4) est négligé car cela signifie seulement que le paramètre λ apporte un changement correspondant et le vecteur final est le même. Puisque la fonction objective $\xi(w)$ est différentiable, sa dérivée par rapport à w peut être calculée comme suit:

$$\frac{\partial \xi(w)}{\partial w_l} = \sum_i \sum_j y_{ij} \left[\frac{2}{\sigma} p_{ij} (\sum_{k \neq i} p_{ik} |x_{il} - x_{kl}| - |x_{il} - x_{jl}| w_l) - 2\lambda w_l \right] \quad (6)$$

$$= \frac{2}{\sigma} \sum_i (p_i \sum_{k \neq i} p_{ik} |x_{il} - x_{kl}| - \sum_j y_{ij} p_{ij} |x_{il} - x_{jl}|) w_l - 2\lambda w_l \quad (7)$$

$$= 2 \left(\frac{1}{\sigma} \sum_i (p_i \sum_{j \neq i} p_{ij} |x_{il} - x_{jl}| - \sum_j y_{ij} p_{ij} |x_{il} - x_{jl}|) - \lambda \right) w_l \quad (8)$$

L'utilisation de la dérivée ci-dessus conduit à l'équation de gradient. Le pseudo code de l'algorithme de NCA est donné ci-dessous [29].

II.3.1.2. Algorithme de NCA

Algorithme : calcul du NCA

1 : **procédure** NCFS ($T, \alpha, \sigma, \lambda, \eta$) où T : ensemble d'apprentissage, α : longueur de l'étape initiale, σ : largeur du noyau, λ : paramètre de régularisation, η : petite constante positive;

2 : Initialisation: $\mathbf{w}^{(0)} = (\mathbf{1}, \mathbf{1}, \dots, \mathbf{1}), \epsilon^{(0)} = -\infty, \mathbf{t} = \mathbf{0}$

3 : **répéter**

4 : **for** $i = 1, \dots, N$ **do**

5 : Calculer \mathbf{p}_{ij} et \mathbf{p}_i en utilisant $\mathbf{w}^{(t)}$ selon les équations (2) et (3)

6 : **for** $l = 1, \dots, d$ **do**

7 : $\Delta_t = 2 \left(\frac{1}{\sigma} \sum_i (\mathbf{p}_i \sum_{j \neq i} \mathbf{p}_{ij} |x_{il} - x_{jl}| - \sum_j \mathbf{y}_{ij} \mathbf{p}_{ij} |x_{il} - x_{jl}|) - \lambda \right) \mathbf{w}_l^{(t)}$

8 : $\mathbf{t} = \mathbf{t} + \mathbf{1}$

9 : $\mathbf{w}^{(t)} = \mathbf{w}^{(t-1)} + \alpha \Delta$

10 : $\epsilon^{(t)} = \xi(\mathbf{w}^{(t-1)})$

11 : **end for**

12 : **end for**

11 : **if** $\epsilon^{(t)} > \epsilon^{(t-1)}$ **then**

12 : $\alpha = 1.01\alpha$

13 : **else**

14 : $\alpha = 0.4\alpha$

15 : **jusqu'à** $|\epsilon^{(t)} - \epsilon^{(t-1)}| < \eta$

16 : $\mathbf{w} = \mathbf{w}^{(t)}$

17 : **revenir** \mathbf{w}

II.4. CLASSIFICATION

La classification consiste à regrouper, dans une même classe de caractère, tous les objets similaires sur la base d'une ressemblance de propriétés (caractéristiques). La règle décision de l'appartenance d'un objet ou d'une forme à une classe donnée est établie à travers un modèle décisionnel. Ainsi, à partir d'un espace multidimensionnel représentatif de l'image (espace de caractéristiques), une nouvelle représentation de la

donnée acquise par exemple par un scanner, désignée par «étiquette» est obtenue et sera évaluée par rapport à une base de données test. C'est dans ce cadre de travail que se situe ce projet de reconnaissance automatique de chiffres manuscrits. Les approches de classification, selon la façon de faire l'apprentissage, peuvent être supervisées ou non supervisées [30].

II.4.1. Classification non supervisée

Dans les méthodes non supervisées, on ne dispose pas de l'ensemble des échantillons d'apprentissage. Le nombre de classes et les règles d'affectation à ces classes doivent être établis seulement à partir d'observations, sans faire référence à une base d'apprentissage. Le groupement des individus est réalisé sur la base de similarités et il est généralement conditionné par le choix du nombre de classes. L'utilisateur n'intervient qu'une fois la classification effectuée pour interpréter le contenu des classes, sans faire appel à d'autres hypothèses sur les images ou sur les classes. Notons que ces techniques sont communément utilisées dans un but exploratoire [30]. On peut distinguer deux types de méthodes de classification non supervisée : les méthodes hiérarchiques et les méthodes non hiérarchiques. Ces dernières sont les plus connues et découlent de l'algorithme de base Iso data [31]. Parmi celles-ci, on cite : la méthode des centres mobiles, la méthode des nuées dynamiques, la méthode des K-means (K-moyennes ou centres mobiles) et la méthode de la logique floue (FCM).

II.4.2. Classification supervisée

Dans les méthodes supervisées, on connaît à partir d'une information a priori sur les données à classifier, le nombre de classes possibles. Une base d'apprentissage est alors construite par la sélection d'un certain nombre d'échantillons de chaque classe. Le problème est d'associer à tout nouvel objet sa classe la plus adaptée, en se servant des échantillons de la base d'apprentissage.

La phase d'apprentissage supervisé consiste à déterminer les règles de décision à partir des exemples de la base d'apprentissage. Pour un ensemble d'apprentissage donné, on peut construire les frontières de décision de deux manières différentes. La première solution consiste à générer les frontières implicitement à partir des distributions de probabilité de chaque classe (**approches par modélisation** : Maximum de Vraisemblance, Modèle de Markov caché, K plus proches voisins). Le deuxième

type d'approche consiste à estimer explicitement les frontières de décision entre les classes (**approches par séparation** : Réseaux de Neurones, Machines à Vecteurs de Support) [32].

II.4.3. Classification par séparation

Les méthodes de classification par séparation visent à construire directement des frontières de décision par la minimisation d'un critère entre les sorties réelles et escomptées du classifieur. Le critère d'erreur choisi est souvent le taux de classification [33]. Parmi les classifieurs qui ont été développés dans ce contexte, nous citons par exemple les réseaux de neurones et les machines à vecteurs de support.

II.4.3.1. Réseaux de neurones

Un réseau de neurones est un assemblage de neurones connectés entre eux. Un réseau réalise une ou plusieurs fonctions algébriques de ses entrées, par composition des fonctions réalisées par chacun des neurones. La capacité de traitement de ce réseau est stockée sous forme de poids d'interconnexions obtenus par un processus d'apprentissage à partir d'un ensemble d'exemples d'apprentissage. Il arrive souvent que les exemples de la base d'apprentissage comportent des valeurs approximatives ou bruitées. Si on oblige le réseau à répondre de façon quasi parfaite relativement à ces exemples, on peut obtenir un réseau qui est biaisé par des valeurs erronées [32].

II.4.3.1.1. Le neurone formel

Le neurone formel est le modèle mathématique du neurone biologique. Il fait la somme pondérée des ses entrées, suivie d'une non linéarité (élément de décision pour les classifieurs) appelée fonction d'activation ou fonction de seuil. Les entrées d'un neurone sont soit des entrées externes, soit des sorties d'autres neurones [37]. Le schéma d'un neurone formel est donné par la figure II.3.

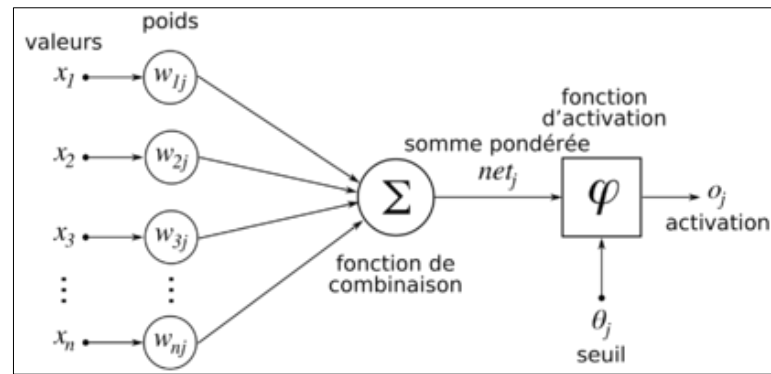


Fig. II.3.Schéma d'un neurone formel.

II.4.3.1.2. Principe de fonctionnement

Pour un neurone formel possédant n entrées, le neurone effectue la somme pondérée suivant :

$$y = \sum_{i=1}^n w_i x_i \quad (9)$$

Puis « active » sa sortie grâce à une fonction non linéaire :

$$z = f(y) = (\sum_{i=1}^n w_i x_i) \quad (10)$$

Le choix de la fonction d'activation dépend de l'application [38].

II.4.3.1.3. Fonction d'activation

C'est une fonction présentée généralement par un non linéarité appelée aussi fonction de seuil. Elle permet de définir l'état interne de neurone en fonction de son entrée totale. Toutes les fonctions d'activation utilisées doivent être différentiables, car l'architecture du réseau de neurones l'impose pour que l'apprentissage soit possible [39]. Les fonctions les plus souvent utilisées sont représentées par la figure II.4.

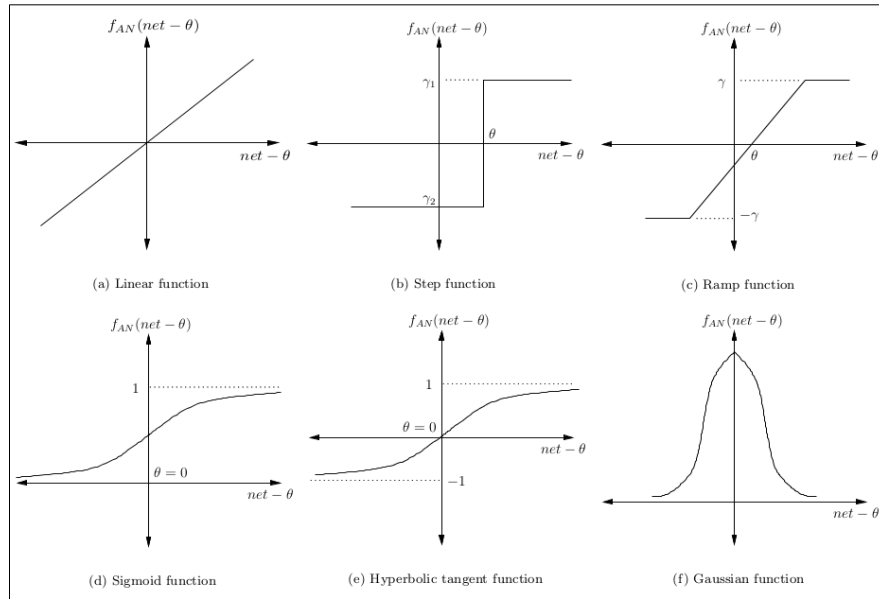


Fig. II.4. Les différentes formes de la fonction d'activation.

II.4.3.1.4. Topologies de réseaux de neurones

Les connexions entre les neurones qui composent le réseau décrivent la topologie du modèle. Elle peut être quelconque, mais le plus souvent il est possible de distinguer une certaine régularité.

Les réseaux multicouche : Ils sont organisés en couches, chaque neurone prend généralement en entrée tous les neurones de la couche inférieure. Ils ne possèdent pas de cycles ni de connexions intra-classe. On définit alors une « couche d'entrée », une « couche de sortie », et n « couches cachées ». Ce type de réseau est très répandu, du fait de son apprentissage aisé [38].

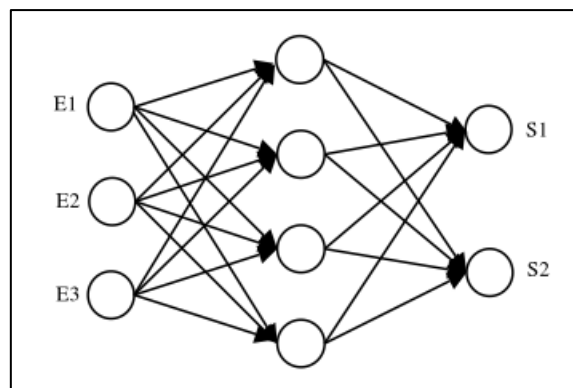


Fig. II.5. Exemple de réseau multicouche a 1 couche cachée, 3 entrées et deux sorties.

Les réseaux a connexions locales : On reprend la même structure en couche que précédemment, mais avec un nombre de connexions limite : un neurone n'est pas forcément connecte a tous les neurones de la couche précédente [38].

Les réseaux à connexion récurrentes : On a toujours une structure en couche, mais avec des retours ou des connexions possibles entre les neurones d'une même couche [38].

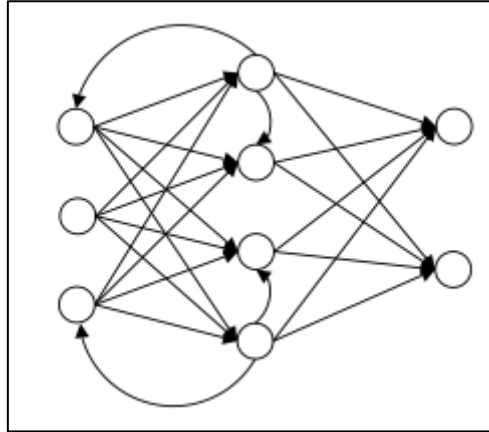


Fig. II.6. Exemple de réseau a connexions récurrentes.

Les réseaux a connexions complètes: tous les neurones sont interconnectés (Cf. Le modèle de Hop-field et la machine de Boltzmann, mis à part l'auto-connexion) [38].

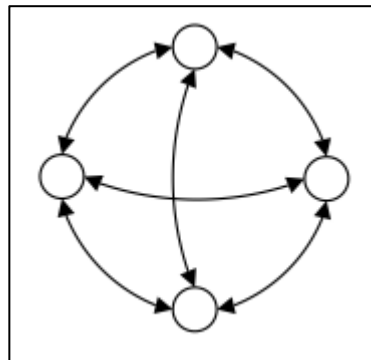


Fig. II.7. Exemple de réseau a connexions complètes.

II.4.3.1.5. Le Perceptron

Présenté originellement par Rosenblatt, en 1958, le perceptron est la forme la plus simple du réseau de neurones. Il permet de classifier correctement des objets

appartenant à deux classes linéairement séparables. L'architecture générale d'un perceptron comprend généralement trois couches principales, La première couche (d'entrée) représente la rétine. Les neurones de la couche suivante sont les cellules d'association, et la couche finale les cellules de décision. La figure II.8 illustre cette architecture [2].

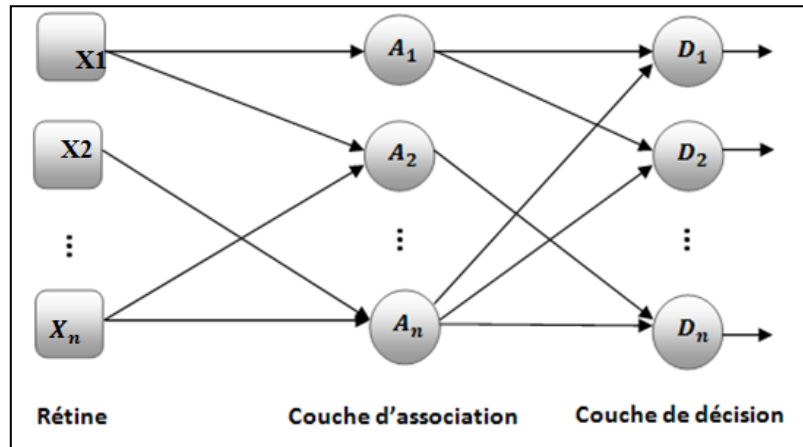


Fig. II.8. Représentation graphique de perceptron.

II.4.3.1.6. Le perceptron multicouche

Le perceptron multicouche (PMC) en anglais « Multi Layer Perceptron : MLP » appartient au réseau multicouche, il ne possède donc pas de boucle de retour, ils sont « Feed-forward ». Les MLP possèdent une fonction d'activation de type sigmoïde ou de heavy-side. Le MLP est une extension multicouche du perceptron, qui est un réseau à une couche, assez limité. Il utilise pour modifier leur poids un algorithme d'apprentissage très répandu car facile à implémenter : la rétro-propagation du gradient, qu'il s'agit toujours de minimiser l'erreur quadratique moyenne (EQM) [38].

Architecture du réseau perceptron multicouches

La mise en cascade de perceptrons conduit à ce qu'on appelle le perceptron multicouches les perceptrons employés ici diffèrent cependant de celui de Rosenblatt par le fait que la non-linéarité utilisée est à présent une fonction continue, d'allure sigmoïdale par exemple, et non plus la fonction de signe. Lorsque le vecteur de caractéristique d'un objet est présenté à l'entrée du réseau, il est communiqué à tous les neurones de la première couche. Les sorties des neurones de cette couche sont alors communiquées aux neurones de la couche suivante. Et ainsi de suite. La dernière couche

du réseau est appelée couche de sortie, les autres étant désignées sous le terme de couches cachées car les valeurs de sortie de leurs neurones ne sont pas accessibles de l'extérieur [2].

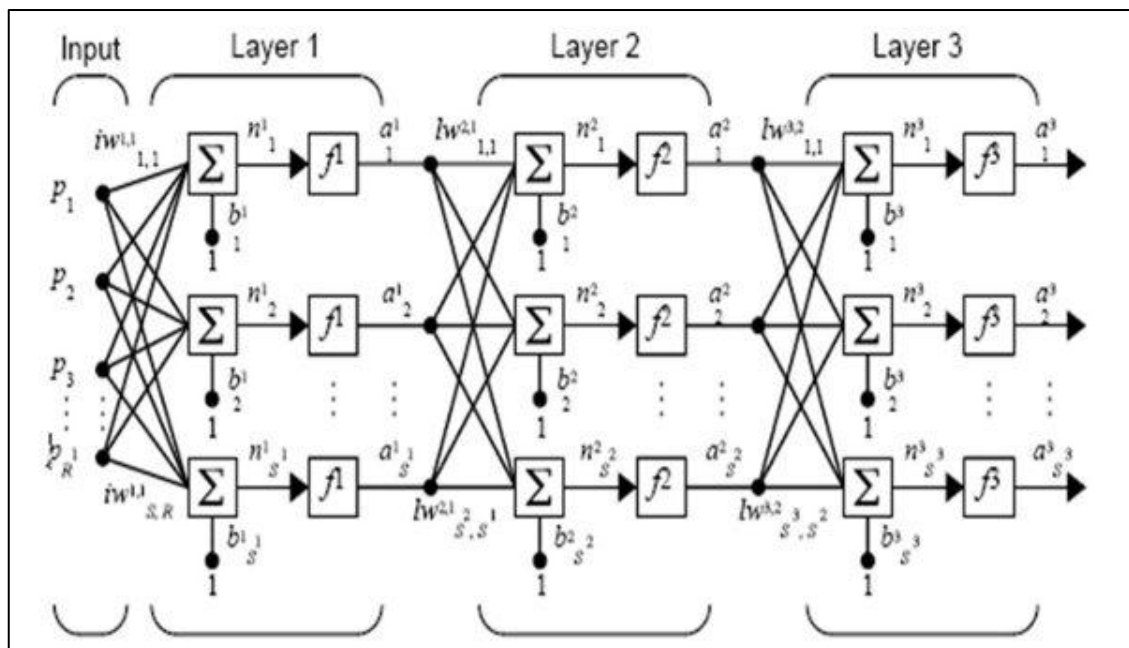


Fig. II.9. Architecture général d'un réseau de neurone MLP.

II.4.3.1.7. Algorithmes d'apprentissage des RNAs

Il existe plusieurs types d'algorithmes d'apprentissage, parmi lesquels nous pouvons citer : la retro-propagation, fast-rétro-propagation, récurrent-retro-propagation et ROM (Random Optimisation- Method).....etc. Dans le présent travail, l'algorithme relatif à la rétro-propagation est particulièrement mis en pratique.

Méthode de rétro-propagation du gradient: La rétro-propagation du gradient est actuellement l'outil le plus utilisé dans le domaine des réseaux de neurones. C'est une technique de calcul des dérivées qui peut être appliquée à n'importe quelle structure des fonctions dérivables.

Principe: La rétro-propagation du gradient consiste à propager « a l'envers » (de la couche de sortie vers la couche d'entrée) l'erreur obtenue sur les exemples de la base d'apprentissage. On utilise pour cela l'erreur quadratique, i.e. la carré de la différence entre ce qu'on obtient et ce qu'on désire.

Si on calcule la dérivée partielle de l'erreur quadratique par rapport aux poids des connexions (d'où le « gradient »), il est possible de déterminer la contribution des poids à l'erreur générale, et de corriger ces poids de manière à se rapprocher du résultat souhaité. La correction se fait par itération en corrigeant plus ou moins fortement les poids par l'intermédiaire d'un coefficient. A l'issue d'un certain nombre d'itérations, lorsque qu'on est satisfait du classement des exemples de notre base d'apprentissage, on fixe les poids qui constituent ainsi des frontières entre les classes [38].

L'algorithme d'apprentissage: Considérons un réseau à une couche cachée.

Le réseau possède :

- Une couche d'entrée à m cellules d'entrées $x_i = e_i$ (Il ne s'agit pas de neurones, ces cellules présentent simplement les entrées e_i au réseau).
- Une couche cachée à n neurones d'activation y_j
- Une couche de sortie à p neurones d'activation z_k
- $n * m$ connexions entre la couche d'entrée et la couche cachée, chacune pondérée par v_{ji}
- $n * p$ connexions entre la couche cachée et la couche de sortie, chacune pondérée par w_{kj}

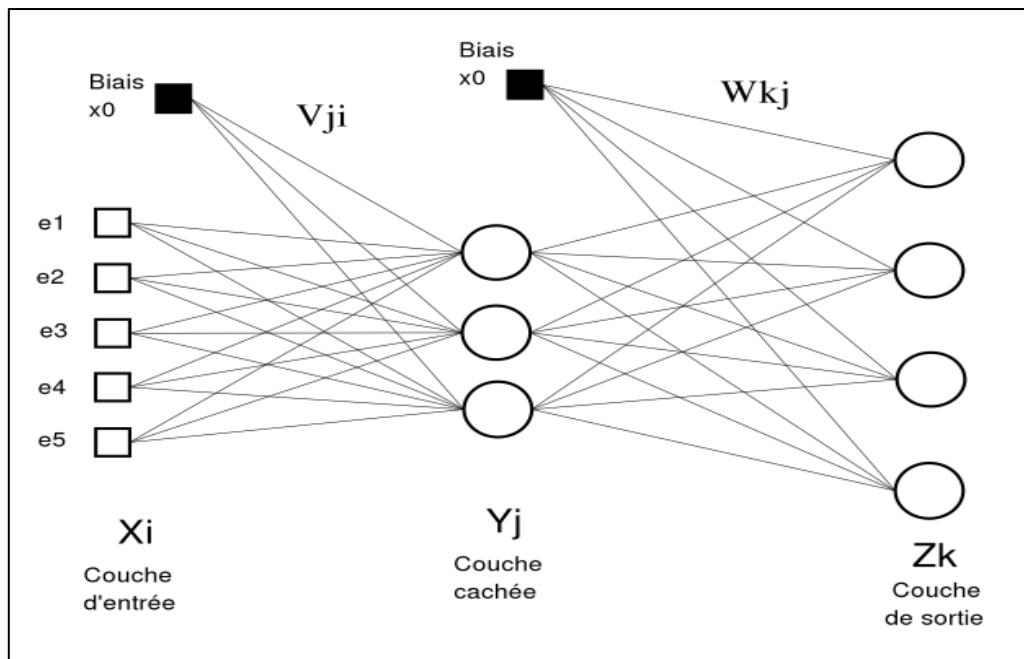


Fig. II.10. Exemple de réseau MLP à une couche cachée avec 5 entrées, 3 neurones dans la couche cachée, et quatre sorties.

ETAPE 1 : Initialisation des poids des connexions Ces poids sont choisis au hasard.

ETAPE 2 : Propagation des entrées les e_i sont présentées à la couche d'entrée : $x_i = e_i$

On propage vers la couche cachée :

$$y_i = f(\sum_{i=1}^m x_i v_{ij} + x_0) \quad (11)$$

puis de la couche cachée vers la couche de sortie :

$$z_k = f(\sum_{j=1}^n y_j w_{kj} + y_0) \quad (12)$$

Les valeurs x_0 et y_0 sont des biais : des scalaires et non des sorties de la couches précédente. La fonction d'activation f choisie est sigmoïde

ETAPE 3 : rétro-propagation de l'erreur Pour chaque exemple de la base d'apprentissage applique en entrée du réseau, on calcule son erreur sur les couches de sorties, c'est à dire la différence entre la sortie désirée s_k et la sortie réelle z_k :

$$E_k = z_k (1 - z_k) (s_k - z_k) \quad (13)$$

On propage cette erreur sur la couche cachée ; l'erreur de chaque neurone de la couche cachée est donnée par :

$$F_j = y_j (1 - y_j) \sum_{k=1}^p w_{kj} E_k \quad (14)$$

ETAPE 4 : Correction des poids des connexions il reste de modifier les poids des connexions :

Entre la couche d'entrée et la couche cachée :

$$\begin{cases} \Delta w_{kj} = n y_j E_k \\ \Delta x_0 = n E_k \end{cases} \quad (15)$$

Entre la couche cachée et la couche de sortie :

$$\begin{cases} \Delta v_{ji} = n x_i F_j \\ \Delta y_0 = n F_j \end{cases} \quad (16)$$

n étant un paramètre qu'il reste à déterminer.

Boucler a l'étape 2 jusqu'a un critère d'arrêt à définir.

Le critère d'arrêt n'est pas forcément évident à définir. Plusieurs solutions : soit on décide d'arrêter quand les erreurs résiduelles sont inférieures à un certain seuil, soit on regarde le comportement du réseau tout au long des itérations sur une base de validation, et on arrête quand commence le phénomène de sur-apprentissage. De même, le paramètre peut se déterminer empiriquement.

II.5. CONCLUSION

Dans ce chapitre, nous avons essayé de donner en bref une description sur le système de reconnaissance des formes et son schéma général. Nous avons aussi présenté le NCA comme une méthode proposée pour la réduction de dimension. Pour la classification, nous avons donné des descriptions sur les différentes méthodes utilisées dans la littérature. Plus de détail a été donné pour les réseaux de neurones artificiels.

L'étude et l'analyse des performances de ces méthodes choisies appliquées au domaine de contrôle et de surveillance des eaux potables, constituent notre principal objectif dans la partie expérimentale. Une étude en simulation ayant pour but d'évaluation des performances de NCA avec les RNAs, fera l'objet du chapitre suivant.

Chapitre III

Évaluation expérimentale

III.1. INTRODUCTION

Ce dernier chapitre est dédié spécifiquement à la mise en œuvre et l'évaluation de notre système proposé qui se base sur l'application de réseaux de neurones artificiels (RNA) sur des données de contrôle et de surveillance des eaux propres. Ces données sont soumis d'une étape de réduction de dimension afin de réduire un peu le nombre des paramètres requis pour le traitement. La technique utilisée dans cette phase est l'analyse en composantes voisines. Également, NCA permet de sélectionner les caractéristiques qui peuvent minimiser la redondance existante entre les variables descripteurs de l'eau. Le protocole expérimental est aussi bien décrit pour bien évaluer les performances de la méthode présentée. Afin d'optimiser les paramètres de chaque module utilisé, nous ferons plusieurs testes. Finalement, une étude récapitulative va clôturer ce chapitre.

III.2. DÉMARCHE PROPOSÉE

Parmi plusieurs méthode de surveillance automatique modernes ont été développées ces dernières années, on trouve les méthodes qui sont basées sur les techniques de l'intelligence artificielles (IA) qui servent comme outil de base pour l'aide à la décision. L'approche de surveillance par reconnaissance de formes est considérée comme une méthode moderne. Cette approche repose sur les méthodes de classification, qui reste le but principale de la reconnaissance de formes. Ces techniques d'IA ont l'avantage de ne pas se baser sur les connaissances a priori des états de fonctionnement, mais plutôt sur une phase d'apprentissage.

Dans notre travail, nous allons utiliser les réseaux de neurones artificiels (RNAs) pour pouvoir prendre une décision sur la qualité des eaux potable d'une manière automatique. Dans cette application, l'approche de surveillance ne s'applique en fait que si on se trouve dans le cas d'un apprentissage supervisé. Nous procédons alors lors d'une étape préliminaire d'apprentissage, à paramétrer le classificateur pour la reconnaissance. A souligner encore une fois que les méthodes de reconnaissance de formes à base d'apprentissage est une étape assez longue et nécessite parfois plus de temps de calcul. Ce module d'apprentissage permet de collecter de manière continue les paramètres relatifs aux différents états de l'eau, ainsi que l'analyse physico-chimique correspondante qui constitue un ensemble de paramètres descripteurs de la qualité de l'eau.

Par ailleurs, afin de trouver les paramètres de l'eau les plus informatifs sur sa qualité, nous utilisons une méthode statistique dite Neighbourhood Components Analysis (NCA) dans une étape préliminaire pour réduire la dimension de la base de données construit des 8 paramètres physico-chimiques de l'eau qui sont : pH, Température (T°), Conductivité (C), Hydrogénocarbonate (B), Magnésium (Mg), Turbidité (TU), FTA et H. La réduction de dimension consiste que l'information essentielle soit conservée et plus facile à analyser que l'ensemble des données d'origine. Dans un autre objectif, et dans le but de construire des capteurs logiciels qui nous fournissent la quantité de quelques paramètres non-mesurables à un prix important, la sélection des paramètres par NCA va nous donner une idée sur l'importance de chaque capteur lorsqu'on veut faire une décision sur sa qualité.

III.3. SYSTÈME DE CONTRÔLE ET DE SURVEILLANCE

Il s'agit dans cette partie de travail d'évaluer les performances de la technique choisie qui est issue, rappelons-le, du domaine de l'intelligence artificielle à savoir, les RNA. Des techniques servant comme outils de base pour l'aide à la décision et présentant une réponse plus élaborée par rapport aux autres techniques se basant sur des données brutes, venant directement des variables de surveillance, ou à partir de données traitées venant des sorties de traitements de bas niveau. Le choix effectué sur la base des résultats obtenus, conduira à l'intégration de la technique sélectionnée au niveau d'un système de surveillance assurant un contrôle permanent de la qualité de l'eau. L'architecture de ce système imaginé est basée sur une approche multi-capteurs et présentée dans la figure III.1. Le processus de contrôle est vu comme un problème de reconnaissance de formes, où les classes correspondent aux différents états de l'eau, et les formes représentent l'ensemble des observations ou mesures des paramètres liés à ses caractéristiques.

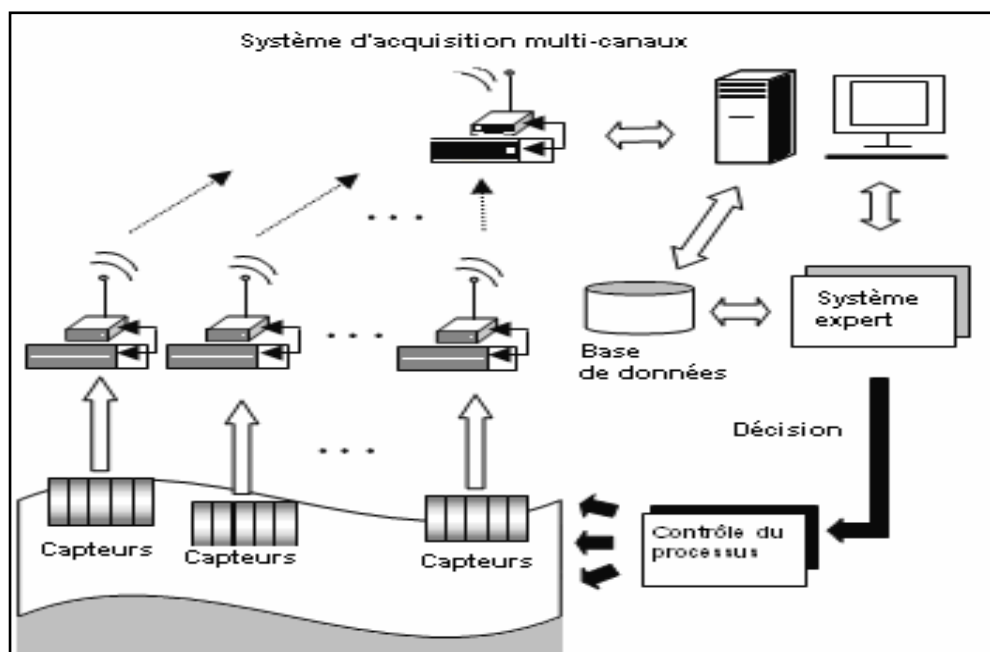


Fig.III.1. Architecture du système de contrôle et de surveillance.

Au niveau du système, on peut supposer que les différents paramètres physico-chimiques utilisées, tels que le pH, la température (T°), la conductivité (C), la turbidité (TU), ... etc. sont transformés en signaux électriques à partir des capteurs physiques, et transmis vers une station de contrôle qui assure l'acquisition, le traitement et l'analyse des données. Ainsi on obtient une nouvelle représentation de l'objet, il existe d'autres modules de traitement pouvant élaborer des représentations ont généralement pour l'objectif de réduire la dimension de la représentation, c'est-à-dire de diminuer le nombre de descripteurs de l'objet à l'entrée et d'élaborer des descripteurs de plus en plus pertinents pour le tâche de discrimination à accomplir telle que notre méthode choisie NCA. La technique de reconnaissance utilisée au niveau du system de décision effectue après chaque acquisition, la classification et la séparation des données en plusieurs classes bien différentes. Une suite d'acquisitions pourrait être envisagée plusieurs fois par jour, sous des conditions prédéfinies. Un module d'apprentissage supervisé par un expert, permet de collecter de manière continue les paramètres relatifs aux différents états de l'eau pour la mise en œuvre d'une base de connaissance complète.

III.3.1. Approche utilisée dans la surveillance

La solution devant être adoptée par les techniques citées ci-dessus au problème de reconnaissance de formes posé, ne s'applique en fait que si on se trouve dans le cas d'un apprentissage supervisé. Nous procédons donc lors d'une étape préliminaire d'apprentissage, à paramétrer le classificateur pour la reconnaissance. L'étape de test ou de reconnaissance proprement dite, s'effectue une fois le modèle statistique établi. Il y a ici tout l'intérêt pour dire que cette approche se caractérise par sa souplesse et sa généralité. A ne souligner toutefois que les méthodes de reconnaissance de formes à base d'apprentissage statistique sont les plus utilisées dans les systèmes de classification à fusion multi sensorielle. En général l'apprentissage est une étape assez longue, et nécessite plus de temps de calcul. Les techniques partagent ce point commun mais diffèrent sur un certain nombre d'autres points. L'étude comparative effectuée dans les paragraphes suivants en fera la différence. Ce critère (temps d'apprentissage) aussi important dans le choix du modèle de reconnaissance, évoque un traitement hors ligne devant être effectué par le système de surveillance. Le déroulement de cette opération en permanence contribue sans doute à enrichir une base de connaissance qu'on veut qu'elle soit la plus complète possible pour le modèle de surveillance implanté. Le système de contrôle doit donc pouvoir marier à la fois une surveillance directe de l'eau et un apprentissage permet de collecter de manière continue les paramètres relatifs aux différents états de l'eau.

III.4. DESCRIPTION DES DONNEES D'ENTREE

Nous cherchons à décider sur la qualité de l'eau à travers ses paramètres descripteurs de la qualité de l'eau brute fournis par les capteurs de mesure de la station. Nous n'avons en fait aucune connaissance a priori sur un type de modèle représentant parfaitement ce procédé, par contre nous pouvons porter notre jugement sur la qualité de cette eau à partir de quelques données descriptives. Il y a quatre paramètres physico-chimiques qui sont souvent utilisés dans plusieurs travaux [50], et qui renseignent sur les dangers majeurs qu'il faut surveiller. Ces paramètres sont résumés comme suit : Conductivité (C), pH, Température (T°), et Turbidité (TU).

L'objectif qui se trouve derrière la collecte des données relatives à ces paramètres est de trouver un modèle de classification permettant de distinguer trois états bien distincts de l'eau (Classe 1, Classe 2 et Classe 3). La qualité de cette eau reflétée par sa

potabilité repose en fait sur une corrélation qui ne peut être identifiée que statistiquement. Des données descriptives expérimentales recueillies sur une période de trois ans (2009-2011) à partir de la station de Tilesdit (Bouira – Algérie) comme étant une zone d'étude dans ce travail qui pourraient atteindre cet objectif. A noter que la turbidité et le pH sont fortement dépendants des phénomènes saisonniers. Il y a donc intérêt de disposer d'au moins une année pour archiver des données afin de déterminer une base de connaissance assez complète capable de fonctionner normalement. D'où la nécessité d'une base de connaissance riche en informations exigeant d'abord une collecte des données sur une longue période, et la présence d'un expert.

Dans un but de validation, deux bases de données réelles sont utilisées. Les différents paramètres descripteurs de l'état de l'eau sont organisés selon les normes recommandées par les pouvoirs publics. Une classification multi-classe comportant trois classes de potabilité différentes (classe 1, 2 et 3) est appliquée sur les données réelles collectées à partir de la station de notre zone d'étude.

III.4.1 Les Bases de données

L'information la plus importante pour ces modèles d'apprentissage statistique est la base d'entraînement. Deux bases de données réelles devraient être utilisées dans ce cas, la première base est constituée de 400 échantillons contient des mesures réelles mais aussi d'analyses chimiques et physiques effectuées hors ligne qui constituent un ensemble de 4 descripteurs (4D) de la qualité de l'eau brute. La deuxième est constituée de 123 échantillons et constituent un ensemble de 8 descripteurs (8D). Chaque échantillon de la deux base possède une classe d'appartenance (3 classe est prévu ; classification multi-classes). Encore une fois, les base sont constituons respectivement de 4 et de 8 paramètres physico-chimiques de l'eau qui sont : pH, Température (T°), Conductivité (C), Hydrogénocarbonate (B), Magnésium (Mg), Turbidité (TU), FTA et H. les 4 premier paramètres sont réserver a la base de 400 échantillons.

III.5. PARTIE EXPÉRIMENTALE

Dans cette partie, nous faisons plusieurs tests afin d'évaluer les performances de notre système proposé. Nous commençons par l'étude et l'évaluation de NCA sur la sélection des paramètres. Ensuite, nous testons les RNAs pour la classification de la qualité des eaux. Il est très important de noter que dans un chaque test, nous utilisons la

technique k-fold-cross-validation pour absorber le problème de l'initialisation aléatoire dans les RNAs et également le problème des données déséquilibrées dans la NCA.

III.5.1. Évaluation de NCA

a. Sur la base 4D:

Dans cette expérience, nous varions le nombre de partition de cross-validation de $k=2$ jusqu'à $k=10$, avec différentes combinaisons de paramètres de l'eau. La figure III.2 et les tableaux (III.1 - 4) montrent les résultats obtenus sur les deux bases de données.

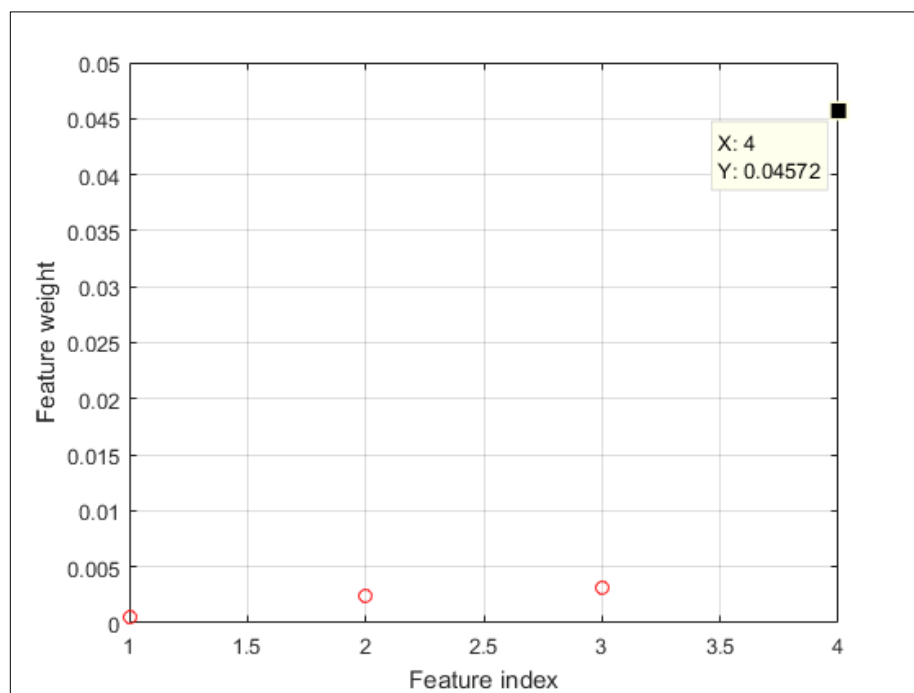


Fig.III.2. Poids de chaque caractéristique en utilisant NCA sur la base 4D.

La figure III.2 représente l'importance (le poids) de chaque caractéristique (paramètres) après l'exécution de la NCA sur la base 4D. Nous remarquons que le 4eme paramètre a une valeur de poids plus élevée par rapport à les autres. Ça vaut dire que celui-ci est le plus représentatif parmi les autres. Afin de vérifier la robustesse de NCA pour sélectionner les paramètres les plus informatifs sur la qualité des eaux, nous avons fait plusieurs tests sur la base 4D en introduisant différentes combinaisons de caractéristiques avec différents nombre de k-fold-cross-validation.

Taux d' apprentissage (%)									
K-fold	2	3	4	5	6	7	8	9	10
TU	99,69	99,69	99,75	99,63	98,81	99,69	99,63	99,69	99,63
TU+T	99,69	99,81	99,75	95,69	96,87	99,75	99,75	99,81	99,88
TU+T+C	99,81	99,69	98,62	98,81	97,75	99,81	98,87	99,69	99,75
TU+T+C+PH	99,69	99,56	99,56	99,88	99,69	99,94	99,56	99,5	99,69

Tab.III.1. Résultats d' apprentissage sur la base 4D en utilisant k-fold-cross-validation avec 4 combinaisons des descripteurs.

Taux de test (%)									
K-fold	2	3	4	5	6	7	8	9	10
TU	99,13	99,25	99,13	99,13	99,13	99,25	99,25	99,13	99,5
TU+T	98,75	99,13	99	94,88	95,88	98,63	99,25	98,75	98,75
TU+T+C	98,5	98,62	98,38	96,87	96,37	98,25	98	98,87	97,75
TU+T+C+PH	98,75	99	98,62	98,75	98,37	98,13	99	98,12	98,62

Tab.III.2. Résultats de test sur la base 4D en utilisant k-fold-cross-validation avec 4 combinaisons des descripteurs.

D'après les résultats reportés dans ces tableaux, nous remarquons que le 4ieme paramètre (TU) donne des meilleurs taux de test et d'apprentissage. ça prouve que TU est plus informatif sur la qualité des eaux. Aussi, les autres combinaisons (TU+T, TU+T+C, TU+T+C+PH) diminue significativement le taux de classification. De l'autre coté, lorsque k= 10 partitions, nous pouvons estimer correctement les poids de chaque paramètre. Cette amélioration est justifiée par la déséquilibration des données.

b. Sur la base 8D:

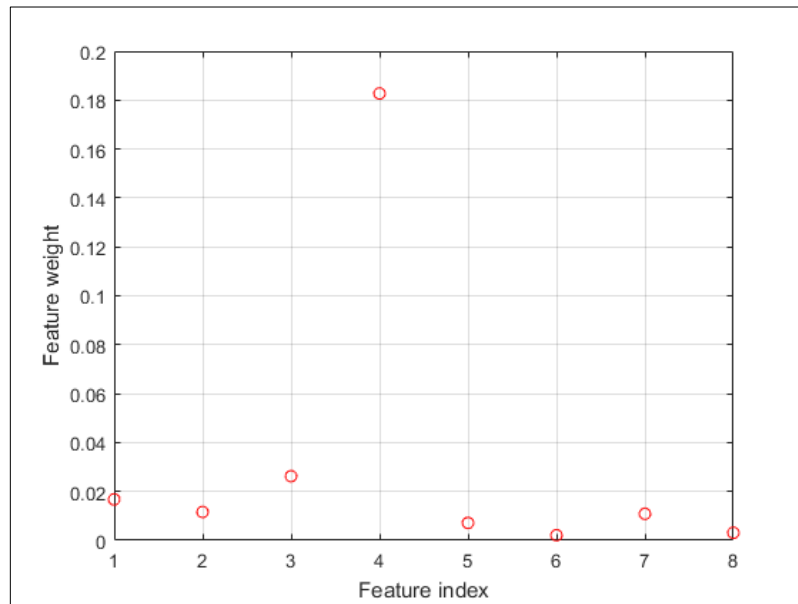


Fig.III.3. Poids de chaque caractéristique en utilisant NCA sur la base 8D.

Taux d'apprentissage (%)								
K-fold	2	3	4	5	6	7	8	9
TU	100	100	100	100	100	100	100	100
TU+T	99,18	100	100	100	100	100	100	100
TU+T+PH	100	100	100	100	100	100	100	100
TU+T+PH+C	100	100	100	100	98,78	98,77	100	100
TU+T+PH+C+H	97,15	100	100	99,59	100	98,77	100	100
TU+T+PH+C+H+Mg	99,59	99,59	100	100	97,53	100	100	99,18
TU+T+PH+C+H+Mg+FTA	100	100	95,47	98,35	98,37	97,95	97,95	100
Tous les paramètres	99,59	100	95,91	98,36	95,06	100	97,55	100

Tab.III.3. Résultats d'apprentissage sur la base 8D en utilisant k-fold-cross-validation avec 8 combinaisons des descripteurs.

Taux de test (%)								
K-fold	2	3	4	5	6	7	8	9
TU	97,56	97,56	98,35	98,35	99,17	98,35	97,54	97,56
TU+T	96,75	97,54	98,37	98,35	96,71	97,52	98,35	99,17
TU+T+PH	93,46	95,87	94,27	97,5	95,89	98,37	95,08	95,12
TU+T+PH+C	95,89	96,57	91,81	93,48	91,81	91,02	96,75	94,29
TU+T+PH+C+H	90,98	95,89	91,85	93,43	92,56	91,06	95,91	85,98
TU+T+PH+C+H+Mg	91,02	86,08	83,7	84,47	77,95	88,62	91,02	85,96
TU+T+PH+C+H+Mg+FTA	86,83	94,27	86,06	86,91	76,16	79,45	95,91	91,04
Tous les paramètres	75,28	81,26	81,95	77,97	77,91	86,81	68,9	63,88

Tab.III.4. Résultats de test sur la base 8D en utilisant k-fold-cross-validation avec 8 combinaisons des descripteurs.

Nous avons fait la même chose pour la base 8D, où la Figure III.3 représente la valeur de poids en fonction des caractéristiques dans le cas de la base 8D. Le 4^{ème} paramètre reste le plus représentatif parmi les autres avec un poids très élevée.

Nous remarquons que les résultats est presque similaire. Pour 10-fold, nous avons le meilleur résultat c.-à-d. Un taux d'apprentissage de 100% et de test de 99.17%. Par ailleurs, la combinaison des autres paramètres diminue le taux d'apprentissage ainsi que le taux de test.

III.5.2. Évaluation des RNAs

Dans cette partie de test, nous appliquons la technique k-fold-cross-validation pour avoir le taux moyen de bonne classification sur les deux bases de données.

a. Sur la base 4D:

K-fold	2	3	4	5
TU	99,63	99,62	99,67	99,66
TU+T	99,63	97,56	99,17	98,41
TU+T+C	97,12	99,81	99,54	99,63
TU+T+C+PH	99,75	98,56	98,54	99,47

Tab.III.5. Résultats d'apprentissage de RNA sur la base 4D en utilisant k-fold-cross-validation avec différentes combinaisons des descripteurs.

K-fold	2	3	4	5
TU	99	99,25	99,25	99,5
TU+T	99,25	97,75	98,88	97,62
TU+T+C	94,87	98	98,63	98,75
TU+T+C+PH	98,88	97,25	97,88	98,75

Tab.III.6. Résultats de test de RNA sur la base 4D en utilisant k-fold-cross-validation avec différentes combinaisons des descripteurs.

Les taux de classification obtenus pour l'apprentissage et le test de RNA sont montrés dans les tableaux **Tab.III.5.** et **Tab.III.6.** Quand nous avons utilisé 5-fold-cross-validation, nous avons trouvé un taux de 99.66% pour l'apprentissage et 99.5% pour le test. Ce phénomène peut être expliqué par l'efficacité de l'utilisation k-fold-cross-validation dans le calcul des performances de RNA.

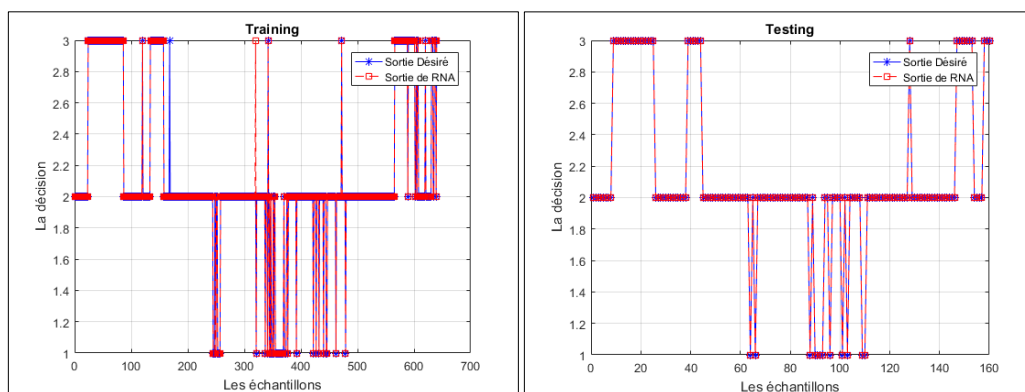


Fig.III.4. La correspondance entre la sortie désirée et de la sortie de RNA dans le cas d'apprentissage et de test sur la base 4D.

La figure ci-dessus montre la correspondance entre les sorties désirées et les sorties calculées via RNA. Nous remarquons que les deux sorties sont presque superposables, ce qui montre que notre le système classe très bien nos données en utilisant les réseaux de neurone artificiels.

b. Sur la base 8D:

Taux d'apprentissage (%)

K-fold	2	3	4	5
TU	100	100	100	100
TU+T	100	100	100	100
TU+T+PH	100	100	99,73	100
TU+T+PH+C	99,18	99,59	100	100
TU+T+PH+C+H	100	98,77	99,73	99,59
TU+T+PH+C+H+Mg	100	100	98,35	98,98
TU+T+PH+C+H+Mg+FTA	100	100	99,73	100
Tous les paramètres	98,36	100	96,2	98,14

Tab.III.7. Résultats d'apprentissage de RNA sur la base 8D en utilisant k-fold-cross-validation avec différentes combinaisons des descripteurs.

Taux de test (%)

K-fold	2	3	4	5
TU	96,72	99,19	98,36	99,2
TU+T	94,26	90,22	98,36	97,53
TU+T+PH	98,36	95,89	96,75	95,93
TU+T+PH+C	85,25	93,46	91,77	96,73
TU+T+PH+C+H	81,15	78,82	92,61	91,87
TU+T+PH+C+H+Mg	81,97	78,84	86,21	84,43
TU+T+PH+C+H+Mg+FTA	75,14	87,7	83,58	84,4
Tous les paramètres	66,39	84,49	85,32	91,13

Tab.III.8. Résultats de test de RNA sur la base 8D en utilisant k-fold-cross-validation avec différentes combinaisons des descripteurs.

D'après les résultats reportés dans les tableaux Tab.III.7. et Tab.III.8. Nous avons remarqué une relation entre le nombre de k-fold et l'amélioration du taux de reconnaissance. A titre d'exemple, pour 5-fold nous avons obtenu un taux de reconnaissance de 100% pour l'apprentissage et 99.2% pour le test est c'est le meilleure résultat parmi toutes les combinaisons faites.

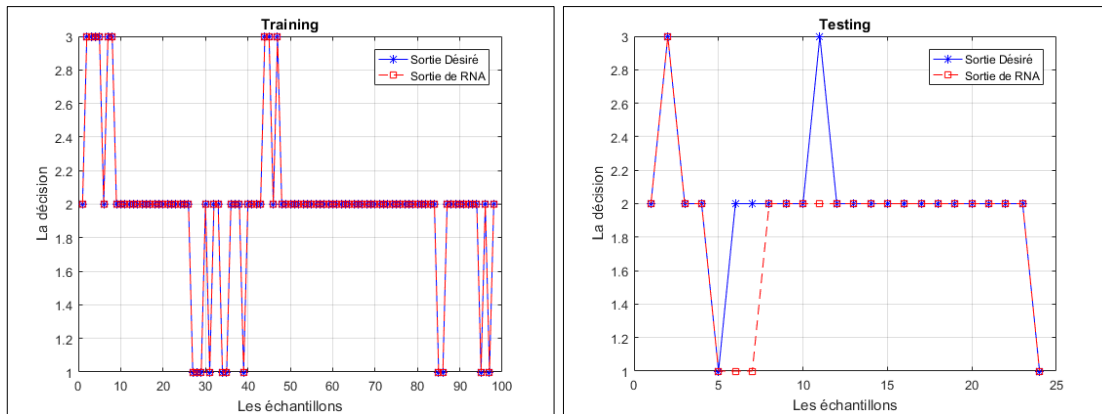


Fig.III.5. La correspondance entre la sortie désirée et de la sortie de RNA dans le cas d'apprentissage et de test sur la base 8D.

En ce qui concerne la classification sur la base de 8D, la figure **Fig.III.4.** montre une correspondance parfaite entre la sortie de RNA et la sortie désire pour l'apprentissage. Dans le test, y avais quelques fausses classifications avec un taux égal à 93.4 %.

III.5.3. L'influence du nombre de neurones dans la couche cachée de RNA

Maintenant, nous voulons savoir l'influence de nombre de neurones dans la couche cachée de RNA sur les performances du système. Pour ce fait, nous avons varié constamment ce nombre tout en gardant le reste fixe. Cette expérience a été appliqué sur les deux bases de données.

a. Sur la base 4D:

Pour connaitre l'effet du nombre de neurone dans la couche cache de RNA sur les résultats de la classification. Nous avons fait un test pour savoir combien de neurone dans la couche cachée faut-il pour une meilleure performance. Nous avons varié leur nombre de 4 neurones jusqu'a 15, les résultats obtenus sont reportés dans la figure ci-dessous.

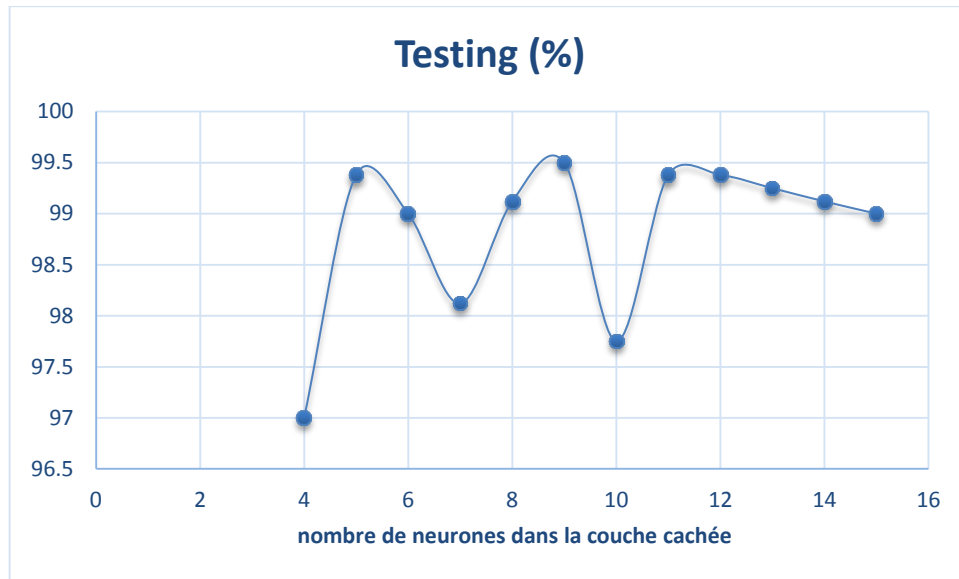


Fig.III.6. Taux de classification (test) en fonction de nombre de neurones dans la couche cachée sur la base 4D.

La figure III.6 montre la variation de taux de test en fonction de nombre de neurones dans la couche cachée sur la base 4D. Il est clair que le taux de classification des eaux est maximal (99.28% pour l'apprentissage et 99.5% pour le test) avec 9 neurones dans la couche cachée de notre RNA.

b. Sur la base 8D:

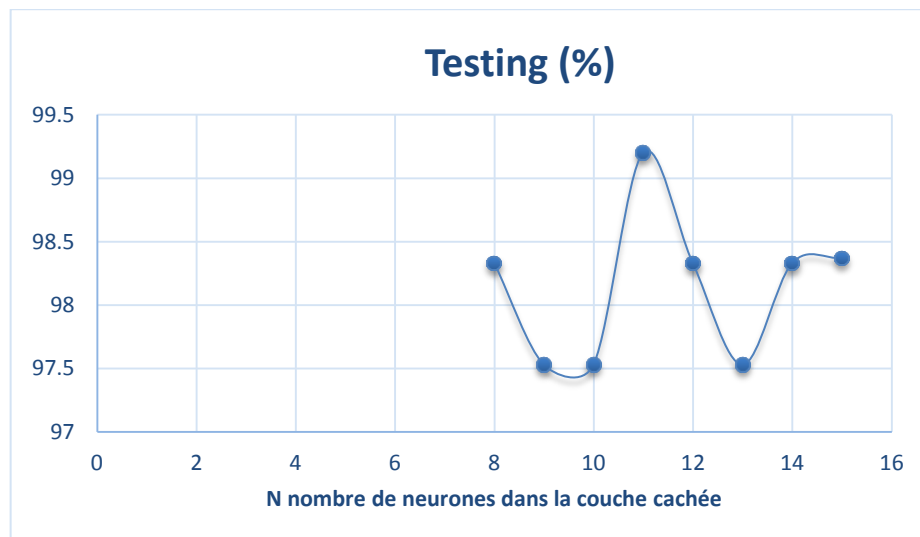


Fig.III.7. Courbe montre le taux de reconnaissance de l'RNA en fonction de nombre de neurones dans la couche cachée (8D).

Pour le cas de la base 8D, le nombre de neurones dans la couche cachée qui nous donne le meilleur résultat est 11 neurones avec un taux de classification parfaite en apprentissage (100%) et 99.2% pour le test.

III.5.4. Le choix de la fonction d'activation pour le RNA

La fonction d'activation joue un rôle très important dans la décision de RNA ce qui influence directement sur les résultats de la classification. A ce fin, nous avons fait une expérience avec plusieurs combinaisons des fonctions d'activation de couche cacher et la couche de sorties afin de trouver la meilleure combinaison.

a. Sur la base 4D

Taux d'apprentissage (%)

Fonction d'activation	{'tansig' 'purelin'}	{'logsig' 'purelin'}	{'hardlim' 'purelin'}	{'purelin' 'purelin'}
TU	99,66	99,66	87,97	84,75
TU+T	99,22	99,34	84,03	86,47
TU+T+C	99,63	99,53	81,78	87,75
TU+T+C+PH	99,75	98,5	75,94	87,72

Tab.III.9. Taux d'apprentissage selon la fonction d'activation de RNA sur la base 4D.

Taux de test (%)

Fonction d'activation	{'tansig' 'purelin'}	{'logsig' 'purelin'}	{'hardlim' 'purelin'}	{'purelin' 'purelin'}
TU	99,12	99,12	87,75	84,63
TU+T	98,38	98,5	84,63	86,25
TU+T+C	99,12	98,75	80,5	87,62
TU+T+C+PH	98,75	97,5	75	87,62

Tab.III.10. Taux de test selon la fonction d'activation de RNA sur la base 4D.

Selon les résultats reportés dans les deux tableaux (Tab.III.9. et 10), nous avons trouvé que la combinaison des fonctions ('tansig''purelin') donne les meilleurs performances par rapport les autres combinaisons.

a. Sur la base 8D

Taux d'apprentissage (%)

Fonction d' activation	{'tansig' 'purelin'}	{'logsig' 'purelin'}	{'hardlim' 'purelin'}	{'purelin' 'purelin'}
TU	100	100	90,58	84,83
TU+T	99,79	100	87,1	84,62
TU+T+PH	100	100	83	86,26
TU+T+PH+C	100	100	83,2	86,08
TU+T+PH+C+H	100	99,8	82,99	86,28
TU+T+PH+C+H+Mg	100	99,18	83,81	86,06
TU+T+PH+C+H+Mg+FTA	100	100	83,82	86,49
Tous les paramètres	99,39	99,18	82,39	85,87

Tab.III.11. Taux d'apprentissage selon la fonction d'activation de RNA sur la base 8D.

Taux de test (%)

Fonction d' activation	{'tansig' 'purelin'}	{'logsig' 'purelin'}	{'hardlim' 'purelin'}	{'purelin' 'purelin'}
TU	99,2	98,37	91,73	82,83
TU+T	95,9	97,57	80,3	83,63
TU+T+PH	96,7	97,53	77,83	85,27
TU+T+PH+C	92,67	95,03	79,57	85,3
TU+T+PH+C+H	87,73	84,3	78,73	85,27
TU+T+PH+C+H+Mg	91,67	72,73	81,23	86,13
TU+T+PH+C+H+Mg+FTA	81,3	74,6	80,33	85,23
Tous les paramètres	78,1	73,7	83,63	83,63

Tab.III.12. Taux de test selon la fonction d'activation de RNA sur la base 8D.

Le même travail a été reproduit sur la base 8D. la comparaison entre les résultats obtenus montre que le meilleur taux d'apprentissage (100%) et le meilleur taux de test (99.2 %) a été construit avec la combinaison ('tansig''purelin').

III.6. CONCLUSION

Ce dernier chapitre a fait l'objet d'une étude en simulation concernant la mise en œuvre de la technique statistique NCA permettant de sélectionner les entrées et éliminer les informations redondantes. Il permet aussi de voir l'intérêt de la réduction de dimension, ainsi que la mise en œuvre de la technique d'apprentissage RNA appliquée dans le domaine du contrôle et de surveillance des eaux potables. Cette étude a permis la validation et l'évaluation des performances de chacune des méthodes présentées.

Une étude détaillée a été faite dans le but d'un choix décisif de l'approche proposée. Les paramètres liés au taux de reconnaissance ont été les facteurs pertinents qui ont permis d'évaluer les méthodes étudiées. Les résultats obtenus ont montré que l'approche NCA-RNA donne des performances considérables en matière de classification de la qualité de l'eau propre.

CONCLUSION GENERALE

Le travail présenté dans ce mémoire a été consacré à la mise en œuvre de la technique d'apprentissage RNA de type MLP combinée avec une technique de réduction de dimension (NCA) pour sélectionner les paramètres les plus pertinents (informatifs) à la reconnaissance de formes dans le domaine de contrôle des eaux potables. Cette étude découle des progrès technologiques importants qui ont été enregistrés ces dernières années dans le but et l'intérêt d'une surveillance moderne et efficace de la qualité des eaux propres. A cet effet, notre modeste travail peut être considéré comme une contribution aux solutions proposées, pour résoudre des problèmes d'intérêt stratégique à préoccupation nationale, utilisant des outils modernes à base de techniques avancées.

La technique d'apprentissage pour au niveau du système de décision sur la qualité de l'eau doit pouvoir effectuer la classification et la séparation de ces données en plusieurs classes (3 classes). Le développement de ce système a été précédé d'une analyse statistique (NCA) pour réduire au-maximum le nombre de capteurs utilisés. NCA a permis de déterminer les corrélations existantes entre les variables caractéristiques de l'eau brute puis de ne conserver que les caractéristiques apportant réellement une information pertinente. Notre travail s'est basé sur l'exploitation des données réelles à partir de la station Tilesdit (Bouira-Algérie) comme zone d'étude dans ce travail. Ces données environnementales résident dans le fait qu'on dispose d'un nombre de données de sorties très faible devant un grand nombre d'entrées. Un enjeu majeur du traitement statistique de ces données est l'analyse multi-variée à but décisionnel. D'un point de vue statistique, ce grand nombre de variables devant un petit nombre d'observations rend l'analyse multi-variée difficile. Une façon de contourner ce fléau de la dimension consiste à réduire cette dimension. La classification multi-classes supervisée est vue comme un problème de reconnaissance de formes avec peu d'observations et beaucoup de variables d'entrées dans un système pour l'aide à la décision.

D'après les résultats obtenus, notre travail a permis d'illustrer la pertinence de ces approches proposées lorsqu'elles sont appliquées à l'analyse de deux bases de données réelles. La classification supervisée obtenue en combinant les techniques NCA et les réseaux de neurones artificiels (RNA) de type MLP comme technique de reconnaissance de formes est appliquée au domaine de surveillance des eaux exprimant les variables physico-chimiques indicatrices à permis de montrer l'effet de l'élimination des paramètres redondants sur la classification supervisée. Ces résultats affirment que la technique NCA donne, dans notre application, un taux très élevé par un système multi-capteur opérant avec moins de voies (capteurs physiques) en entrée (c.-à-d moindre coûteux). L'intérêt d'appliquer cette technique dans ce domaine est donc bien justifié.

Plusieurs perspectives peuvent être envisagées, d'une part concernant la méthode de classification, qui peut être complétée avec l'utilisation d'autres algorithmes en parallèle. D'autres part, concernant l'entrée du système en ajoutant des techniques de maximisation de l'information à partir des capteurs intelligents.

REFERENCES

- [1] M. Ladjal, "Contribution au développement de systèmes de surveillance innovants dédiés au contrôle de la qualité des eaux potables", Thèse de Doctorat, Laboratoire LASS, Université de M'sila, 2013.
- [2] M. Saghiour, "Evaluation et détermination des variables d'entrées pour un modèle multi-classe intelligent basée sur la réduction de dimension", Mémoire de Master Université de M'sila, juin 2012.
- [3] M. R. Zemouri, Contribution à la surveillance des systèmes de production à l'aide des réseaux de neurones dynamique, Application à la e-maintenance, Thèse de doctorat, Université de Franche-Comté, 2003.
- [4] R. A. Reyna Rojas, "Conception et intégration VLSI d'un système de vision générique, Application à la détection et la localisation d'objets à l'aide de support vector machines", Thèse de doctorat, Laboratoire LAAS – CNRS, N°02226, Toulouse, France, 2002.
- [5] "Le cycle de l'eau, Observation régionale de l'environnement", Les Agences de l'eau Loire- Bretagne et Adour-Garonne, Poitou-Charentes, France, 2012.
- [6] [Http://www.aquafil.com.au/natural-sources-water/](http://www.aquafil.com.au/natural-sources-water/).
- [7] K.COULIBALY, étude de la qualité physico-chimique et bactériologique de l'eau des puits de certains quartiers du district de Bamako, UNIVERSITÉ DE BAMAKO, Avril 2005.
- [8] [Http://www.english-online.at/geography/water/water-cycle-treatment.htm](http://www.english-online.at/geography/water/water-cycle-treatment.htm).
- [9] C. CARDOT, Les traitements de l'eau. Procédés physico-chimiques et biologiques, Ellipses Edition Marketing S.A, 1999.
- [10] N.VALENTIN Construction d'un capteur logiciel pour le contrôle automatique du procédé de coagulation en traitement d'eau potable, Thèse de doctorat, UTC/Lyonnaise des Eaux/CNRS, 2000.
- [11] H. Hernández, Supervision et diagnostic des procédés de production d'eau potable, Thèse de doctorat, L'INSA de Toulouse 2006.
- [12] CIDF Centre International De Formation. Principes généraux de traitement des eaux, Lyonnaise des Eaux, 2000.

- [13] DEGREMONT, Mémento technique de l'eau : Tome 2. Lavoisier SAS – Lexique technique de l'eau, Paris, dixième édition, 2005.
- [14] https://www.oieau.org/ReFEA/fiches/AnalyseEau/Physico_chimie_PresGen.htm.
- [15] P. Frank, Fault Diagnosis in Dynamic Systems Using Analytical and Knowledge Based Redundancy – A survey and New Results, *Automatica*, Vol. 26, p. 459-474, 1990.
- [16] Q. Zhang, Identification et Surveillance de systèmes Dynamiques, Habilitation à diriger les recherches, Université de Rennes1, Institut de Formation Supérieure en Informatique et en Communication, 1999.
- [17] A.S. Willsky, A Survey of Design Methods for Failure Detection in Dynamic Systems, *Automatica*, Vol. 12, p. 601-611, 1976.
- [18] X. Desforges, Méthodologie de surveillance en fabrication mécanique : application de capteur intelligent à la surveillance d'axe de machine-outil, Thèse de Doctorat, Université de Bordeaux I, 1999.
- [19] R. Zemouri, D. Racoceanu, N. Zerhouni, Réseaux de neurones récurrents à fonctions de base radiales RRFR : Application à la surveillance dynamique, *Revue Systèmes /JESA*, Vol. 37, N°1, pp. 49-81, 2003.
- [20] Nicolas Pallaut, Méthodologie de surveillance dynamique à l'aide des réseaux neuro-flous temporels, Thèse de Doctorat, Université de Franche-Comté de Besançon, France, N° d'ordre : 1135, 2006.
- [21] B. Dubuisson, E. Boutleux, P. Dague, T. Denoeux, E. Didelet, Y. Gandvalet et M. Masson, *Diagnostic, Intelligence artificielle et reconnaissance de formes*, Edition Hermes, France, 2001.
- [22] Daniel Racoceanu, Contribution à la surveillance des systèmes de production en utilisant les techniques de l'intelligence artificielle, Thèse d'HDR, Université de Franche-Comté de Besançon, France, N° d'ordre : 140, 2006.
- [23] L'eau dans un monde qui change 3è Rapport mondial des Nations Unies sur la mise en valeur des ressources en eau.
- [24] Laurence Likforman-Sulem et Elisa Barney-Smith, *Reconnaissance des formes - Théorie et pratique sous Matlab Cours et exercices corrigés*, Ellipses.
- [25] Benahmed, N., 2002. Optimisation de réseaux de neurones pour la reconnaissance de chiffres manuscrits isolés: sélection et pondération des primitives par algorithmes génétiques, Mémoire: Ecole de Technologie Supérieurs : Université du Québec. 124 pages.

- [26] Mébarka Belahcen, Authentification et Identification en Biométrie. Doctoral thesis, Université Mohamed Khider Biskra, 2013.
- [27] Sabra EL FERCHICHI, Sélection et Extraction d'attributs pour les problèmes de classification, Doctoral thesis, Université des Sciences et Technologies Lille 1, 2013.
- [28] Bounneche Meriem Dorsaf, Réduction de données pour le traitement d'images, Mémoire de Magistère, Université Mentouri Constantine, 2009.
- [29] Yang, W., Wang, K., & Zuo, W. (2012). Neighborhood Component Feature Selection for High-Dimensional D
- [30] N. Abbas. Développement de modèles de fusion et de classification contextuelle d'images satellitaires par la théorie de l'évidence et la théorie du raisonnement plausible et paradoxal, Mémoire de magister en électronique, spécialité Traitement du signal et d'images, Université des Sciences et de la Technologie Houari Boumediène, Bab Ezzouar, Alger, Algérie : 2009, 79 p.
- [31] Dunn J. C. "A Fuzzy Relative of the ISODATA Process and its Use in Detecting Compact Well-Separated Clusters", *J. Cybernetics*, vol. 3, No. 3, pp. 32-57, 1973.
- [32] Hamouchene Izem & Oulad dadda Hacene, Combinaison de classifieurs SVMs par la DST et la DSMT pour la reconnaissance des chiffres manuscrits, Mémoire de master, Université des sciences et de la technologie Houari Boumediene, 2011.
- [33] Brik, Y. Reconnaissance des chiffres Manuscrits par les Modèles de Markov Cachés Continus. Mémoire de magister en électronique, spécialité Traitement du signal et d'images, Université des Sciences et de la Technologie Houari Boumediène, Bab Ezzouar, Alger, Algérie, 60 p, 2010.
- [34] Vapnik, V. "The nature of statistical learning theory". Springer Verlag, New York, USA, 1995.
- [35] Zzouz, H et Zerargui, M. Application des SVMs pour la Verification Hors-Ligne Des Signatures Manuscrites. Projet de fin d'études en électronique, option Contrôle et Télécommunication, Université des Sciences et de la Technologie Houari Boumediène, Bab Ezzouar, Alger, Algérie, 47 p, 2010.
- [36] Nemmour, H. Application des Machines à Vecteurs de Support pour la Reconnaissance des Chiffres Manuscrits, Mémoire de magister en électronique,

spécialité Traitement du signal et d'images, Université des Sciences et de la Technologie Houari Boumediène, Bab Ezzouar, Alger, Algérie : 2010, 163 p.

[37] B. Mohamed, Identification et contrôle avec réseaux de neurones, thèse de magistère, Institut d'électronique Sétif, 1993.

[38] Clément Chatelain, les Multi-Layer Perceptron(MLP),14 novembre 2003.

[39] Gilles Zwingelstien, Diagnostic des défaillances, théories et pratique pour les systèmes industriels.

MEMOIRE DE FIN D'ETUDES EN VUE DE L'OBTENTION DU DIPLOME DE MASTER EN GENIE ELECTRONIQUE

Option : ELECTRONIQUE DE SYSTEME EMBARQUE.

Propose et dirige par : Mr. BRIK Youssef et Mr. LADJAL Mohamed.

Étudié par : BACHIRI Mohamed et DAKHANE Mohammed.

Thème : Évaluation et détermination des variables d'entrée pour un modèle multiclasse intelligent dans les procédés de traitement des eaux propres.

Résumé :

L'eau est un élément essentiel dans la vie. Il est donc nécessaire de la contrôler, la surveiller pour la rendre potable. Dans ce cadre des efforts ont été effectués durant ces dernières années, pour développer des méthodes de contrôle et de surveillance automatique de l'eau. Ce travail s'inscrit dans ce courant de recherche, en proposant une nouvelle approche pour le contrôle et la surveillance de la qualité des eaux potables. Ce mémoire aborde le problème de la surveillance des eaux potables en fonction des caractéristiques physicochimiques de l'eau brute à l'aide d'un modèle basé sur les techniques de l'apprentissage statistiques telles que les réseaux de neurones artificiels. la réduction de la dimensionnalité des donnée est la partie principale dans ce travail, elle consiste à retenir les variables les plus représentatives des données observées. Elle peut être utile comme étape préliminaire à tout processus d'analyse et de classification de données à fin de se concentrer sur les variables les plus importantes et réduire le cout d'exécution d'un tel processus. Notre approche par extraction permet non seulement de réduire la dimensionnalité des données mais également de détecter diverse corrélation dans les données observées.

Mots clés : Eau propre, Surveillance, classification, Réduction de dimension, Extraction de données, NCA, RNA, Simulation,

ملخص:

الماء عنصر أساسي في هذه الحياة. لذلك من الضروري التحكم فيها، و مراقبتها لجعلها صالحة للشرب. في هذا السياق، بذلت جهود بحث في السنوات الأخيرة لتطوير طرق المراقبة الآلية للمياه. هذا العمل هو جزء من اتجاه هذا البحث ، الذي يقترح منهجاً جديداً للتحكم في جودة مياه الشرب ومراقبتها. تتناول هذه المذكرة مشكلة مراقبة مياه الشرب وفقاً للخصائص الفيزيائية الكيميائية للمياه الخام باستخدام نموذج يعتمد على تقنيات التعلم الإحصائي مثل الشبكات العصبية الاصطناعية. تقليل أبعاد البيانات هو الجزء الأساسي من هذا العمل، و يتمثل في الاحتفاظ بالمتغيرات الأكثر تمثيلاً للبيانات التي تمت ملاحظتها. يمكن أن تكون مفيدة كخطوة أولية في أي عملية لتحليل البيانات وتصنيفها من أجل التركيز على أهم المتغيرات وتقليل تكلفة تنفيذ مثل هذه العملية. لا يقلل منهجنا الاستخراج من بُعدية البيانات فحسب، بل يكشف أيضاً عن الارتباطات المختلفة في البيانات المرصودة.

كلمات مفتاحية : المياه النقية، المراقبة، التصنيف، تقليل الأبعاد، استخراج البيانات، NCA، RNA، محاكاة