

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE
SCIENTIFIQUE

UNIVERSITE MOULOD MAMMARI DE TIZI-OUZOU



FACULTÉ DE GÉNIE ÉLECTRIQUE ET INFORMATIQUE
DEPARTEMENT D'ÉLECTRONIQUE

Thèse de doctorat 3^{ème} cycle (LMD)

Filière : Génie Electrique

Spécialité : Electronique

Présentée par : **Mr Amar BELGHIT**

Thème :

‘Quantification des précipitations à partir du radiomètre SEVIRI et validation en utilisant les observations de radar météorologique ‘

Thèse soutenue le 10/ 07/ 2024

Devant le Jury composé de :

Mr. LAHDIR Mourad	Professeur à l'UMMTO	Président
Mr. LAZRI Mourad	Professeur à l'UMMTO	Directeur de thèse
Mme.RAHMOUNE-AOUDIA Fazia	Professeur à UAMBejaia	Examineur
Mr. HAMACI Samir	HDR à ECAM-EPMI Cergy Pontoise	Examineur
Mr. MOHIA Yacine	Maître de Conférences (A) à L'UMMTO	Examineur
Mr. AMEUR Soltane	Professeur à l'UMMTO	Invité

Remerciements

Ce travail a été réalisé au laboratoire LAMPA (Laboratoire d'Analyse et de Modélisation des Phénomènes Aléatoires) de l'Université Mouloud MAMMERI de Tizi-Ouzou.

Je tiens à exprimer mes sincères remerciements à mon directeur de thèse Mr. Mourad LAZRI, Professeur à l'Université Mouloud Mammeri de Tizi ousou, pour qui j'exprime ma profonde gratitude pour m'avoir fait confiance en m'encadrant dans cette thèse et de m'avoir fait profiter de ses riches connaissances et savoir dans le domaine, mais aussi pour sa rigueur scientifique, et ses précieux conseils pour finaliser cette thèse, durant laquelle il n'a jamais ménagé ses efforts ni son temps précieux, sa motivation et disponibilité et ses encouragement grâce aux quels j'ai pu mener au bout ce travail .

Je souhaite exprimer ma gratitude toute particulière à Monsieur Mourad LAHDIR Professeur au département d'électronique de l'UMMTO pour avoir accepté de présider le jury de ma soutenance.

J'exprime ma profonde gratitude à Monsieur Samir HAMACI Docteur HDR à ECAM-EPMI de Cergy Pontoise pour avoir accepté de faire partie de mon Jury.

Je tiens à remercier vivement madame RAHMOUNE-AOUDIA Fazia Professeur à l'Université Abderahmane Mira de Béjaia d'avoir accepté de faire partie de mon Jury.

J'adresse également toute ma reconnaissance et remerciements à Monsieur Yacine MOHIA, Maitre de conférences A à l'UMMTO, de m'avoir fait l'honneur d'accepter de participer au Jury de ma soutenance.

Je suis reconnaissant à Monsieur Soltane AMEUR, Professeur à l'UMMTO et Directeur du laboratoire de recherche LAMPA, d'avoir mis ma disposition les moyens nécessaires pour réaliser mon travail dans de bonnes conditions au sein du laboratoire LAMPA.

Je tiens à remercier chaleureusement monsieur Fethi OUALLOUCHE Maitre de conférences et membre du laboratoire LAMPA, pour ces conseils précieux et les discussions que nous avons eu sur le sujet de mon travail de thèse, et je remercie aussi mon ami Ali Hamroune pour son soutiens, toute l'équipe chercheurs et doctorants du laboratoire LAMPA qui ont

contribués de prés ou de loin dans mon parcours vers la finalisation de ce travail, qu'ils trouvent ici l'expression de mes sincères remerciements.

Dédicaces

A ma mère et mon père sans les quels je ne pourrais pas en arriver la et finir ce travail qui a pris tant d'année, merci pour votre soutiens et patience.

A ma sœur Faiza pour son aide et soutiens durant cette période de finalisation de ma thèse, et à son mari Karim et la petite ELOUIZ.

A mes sœurs Amel et Leila et à mon frère Nassim pour leurs soutiens et support.

A la mémoire de ma grand-mère décédée cette année, sans pouvoir être la pour assister à ma soutenance.

A toute ma famille.

A tout mes amis du département électronique et de l'université UMMTO, et a toute personne ayant fait partie de mon parcours de prés ou de loin, et qui a contribué à mon parcours pour mener a terme cette thèse, et tout au long de mon cursus universitaire, par leurs soutiens moral ou aide.

Merci

Amar.B

Acronymes et abréviations

Acronymes et abbreviations:

AdaBoost: Adaptive Boosting.

AdaOvA-SVM: AdaBoost One-versus-All SVM.

ANN: Artificial Neural Networks.

BBT: Balanced Binary Tree.

BTD: Brightness Temperature Difference.

CART: Classification and Regression Trees.

CNN: Convolutionnal Neural Networks.

CS-RADT: Convective/Stratiform Rain Area Delineation Technique.

CTH: Cloud Top Height.

CTT: Cloud Top Temperature.

CWP: Cloud Water Path.

DDAG-SVM: decision directed acyclic graph-SVM.

DMLP: Deep MultiLayer Perceptron.

ECST: Enhanced Convective Stratiform Technique.

ESA: European Space Agency.

EUMETSAT: European Organisation for the Exploitation of Meteorological Satellites.

FAR: False Alarm Ratio.

GCSM : Groupe de Coordination des Satellites Météorologiques.

GERB: Geostationary Earth Radiation Budget.

HRV: High Resolution Visible.

IPS: Instant Precipitation Scene.

IWP: Ice Water Path.

Acronymes et abréviations

JAXA: Japan Aerospace Exploration Agency.

KNN: K-Nearest Neighbours.

LWP: Liquid Water Path.

MAE: Mean Absolute Error.

MBE: Mean Bias Error.

METOP: Meteorological operational satellite.

MFG: Meteosat First Generation

MLP: MultiLayer Perceptron.

MSG: Meteosat Second Generation.

NASA: National Aeronautics and Space Administration.

NIR: Near InfraRed.

NOAA: National Oceanic and Atmospheric Administration.

OMM: Organisation météorologique mondiale.

ONM : Office Nationale de Météorologie.

OSH : Optimal Separating Hyperplane.

OvA: One versus All.

OvO: One versus One.

PERSIANN: Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks.

PMW: Passive Micro Wave.

POD: Probability Of Detection.

POFD: Probability Of False Detection.

RADAR: Radio Detection And Ranging.

Acronymes et abréviations

RBF: Radial Basis Function.

RFT: Random Forest Technique.

RMSE: Root Mean Square Error.

RNN: Recurrent Neural Network.

SART: Support vector machine, Artificial neural network and Random forest Technique.

SBT-SVM: Slant Binary Tree SVM.

SEVIRI: Spinning Enhanced Visible and Infra Red Imager.

SVM: Support Vector Machines.

TAMSAT: Tropical Applications of Meteorology using SATellite data.

WV: Water Vapor.

Sommaire

Introduction	1
--------------------	---

Chapitre I Instruments de mesure et d'observation des précipitations

I.1. Préambule	5
I.2. Précipitations et leurs types.....	6
I.2.1. Précipitations convectives	6
I.2.2. Précipitations stratiformes.....	6
I.3. Instruments de mesure et d'observation des précipitations.....	6
I.3.1. Instruments de mesure au sol.....	7
I.3.2. Instruments de mesures à distance	8
I.3.2.1. Radars météorologiques.....	8
I.3.2.1.1. Mesures par le radar météorologique.....	9
I.3.2.2. Satellites météorologiques	12
I.3.2.2.1. Satellites météorologiques à orbite polaires	14
I.3.2.2.2. Satellites météorologiques géostationnaires	15
I.3.2.2.3. Satellites Météorologiques Météosat	16
I.3.2.2.4. Satellite Météosat deuxième génération.....	18
I.3.2.2.4.1. Le radiomètre SEVIRI	19
I.4. Discussion	29

Chapitre II Etat de l'art sur l'apprentissage automatique (Machine Learning)

II.1. Préambule.....	31
II.2. Définition de l'apprentissage automatique (Machine Learning).....	31
II.3. Application de l'apprentissage automatique.....	33
II.4. Types d'apprentissage automatique	34
II.4.1. Apprentissage automatique supervisé	34
II.4.2. Apprentissage non supervisé.....	35
II.4.3. Apprentissage semi-supervisé.....	35
II.4.4. Apprentissage par renforcement.....	36
II.4.5. Apprentissage profond (Deep Learning).....	37

Sommaire

II.4.5.1. Réseaux de neurones convolutifs(CNN).....	38
II.4.5.2. Réseaux de neurones récurrents (RNN).....	39
II.4.5.3.Auto Encoders (AE).....	39
II.5. Choix du model apprentissage automatique et modélisation du problème.....	40
II.5.1. Choix du model d'apprentissage automatique.....	41
II.5.2. Modélisation d'un problème.....	41
II.5.3. Cas d'apprentissage automatique supervisé.....	42
II.5.4. Décision dans le cas de classification.....	43
II.5.4.1. Classification Une-contre-toutes (ONE vs ALL).....	44
II.5.4.2. Classification Une-contre-Une (ONE vs ONE).....	44
II.6. Préparation de données et fonctionnement en cas de la classification.....	45
II.6.1.Etapes de préparations des données.....	45
II.6.1.1. Acquisition des données.....	45
II.6.1.2. Prétraitement.....	45
II.6.1.3. Conversion.....	46
II.6.1.4. Post-traitement.....	47
II.6.2. Etapes de fonctionnement de l'apprentissage automatique.....	47
II.6.2.1. Apprentissage.....	47
II.6.2.2.Etape de validation.....	47
II.6.2.3. Etape de test.....	48
II.6.3. Validation croisée.....	48
II.7.Notions de généralisation et sur-apprentissage.....	49
II.7.1.Généralisation.....	49
II.7.2. Sur-apprentissage.....	50
II.7.3. Sous-apprentissage.....	50
II.8. Quelques modèles d'apprentissage automatique.....	51
II.8.1.Réseaux de neurones artificiels.....	51
II.8.2. Algorithmes des K plus proches voisins (KNN).....	54
II.8.3. Algorithme Machines à Vecteurs Support (SVM).....	56

Sommaire

II.8.4. Arbres de décision.....	60
II.8.5. Méthodes d'ensemble (ensemblistes).....	60
II.8.5.1. Méthodes d'ensemble parallèles.....	61
II.8.5.2. Méthodes d'ensemble séquentielles.....	61
II.9. Discussion.....	63

CHAPITRE III Méthodes d'estimation des précipitations par satellite

III.1. Préambule.....	64
III.2. Méthodes d'estimation des précipitations par satellite.....	65
III.2.1. Méthodes infrarouge (IR).....	65
III.2.2. Méthodes multi spectrales.....	67
III.2.3. Méthodes basées sur l'apprentissage automatique.....	70
III.2.3.1. Réseaux de neurones artificiels.....	71
III.2.3.2. Arbre de décision.....	72
III.2.3.3. Méthode SVM (Support Vector Machines).....	72
III.2.3.4. Méthodes d'ensembles.....	73
III.2.3.5. Méthode KNN (K nearest-neighbours).....	75
III.2.3.6. Méthodes à multi-classifieurs.....	75
III.2.4. Méthodes apprentissage profond.....	78
III.3. Méthodes Micro-ondes.....	80
III.4. Méthodes combinant les données MO/multispectrales.....	81
III.5. Discussion.....	82

CHAPITRE IV Estimation des précipitations par la méthode AdaBoost-OvA-SVM

IV.1. Préambule.....	84
IV.2. Zone d'étude et données.....	85
IV.2.1. Zone d'étude.....	85
IV.2.2. Données.....	86
IV.2.2.1. Data MSG.....	87

Sommaire

IV.2.2.2.Données Radar.....	87
IV.2.2.3.Données pluviographiques.....	88
IV.2.2.4. Correspondance entre les données satellite et les données Radar.....	89
IV.2.2.5.Sélection des paramètres d'entrées.....	89
IV. 3. Contexte scientifique de l'étude.....	92
IV.4. Méthodologie.....	94
IV.4.1. Concept mathématique d'AdaBoost.....	94
IV.4.2. Concepts mathématiques de Support Vector Machines (SVM).....	96
IV.4.3. Elaboration des modèles de classification avec AdaBoost.....	102
IV.5.Tests et résultats des classifications.....	106
IV.5.1.Application sur une scène instantanée.....	106
IV.5.2.Application sur toutes les scènes.....	110
IV.5.3. Inter-comparaison.....	113
IV.6. Estimation des taux de précipitations.....	116
IV.7.Discussion.....	121
Conclusion.....	122
Annexes.....	124
Bibliographie.	

Liste des figures

Figure I.1 : Système mondiale des satellites d'observation météorologiques et leurs distributions (Eumetsat,2005).....	13
Figure I.2 : Orbite d'un satellite à défilement.....	15
Figure I.3 : Zones couvertes par les différents satellites météorologiques géostationnaires (source EUMETSAT [Eumet]).....	16
Figure I.4 : zone de couverture des satellites METEOSAT.....	17
Figure I.5 : Concept de scan du satellite METEOSAT (source EUMETSAT [Eumet]).....	19
Figure I.6 : comparaison entre les satellites MFG et MSG de METEOSAT.....	21
Figure I.7 : images acquises des différents canaux de MSG-SEVIRI le 01 aout 2006 à 13 :00 UTC.....	23
Figure I.8 : Principe d'acquisition des images des satellites Météosat.....	27
Figure I.9 : Caractéristiques des images MSG.....	28
Figure I.10 La résolution radiométrique.....	29
FigureII.1 la relation entre l'Intelligence Artificielle, le Machine Learning et le Deep Learning.....	33
Figure II.2 Grandes classes ou types de l'apprentissage automatique.....	36
Figure II.3 Schéma de l'architecture de base d'un réseau CNN.....	38
Figure II.4 Connexions récurrentes dans un réseau RNN.....	39
Figure II.5 Schéma de l'architecture de l'Auto-Encodeur.....	40
Figure II.6 Illustration de modèle en sous-apprentissage et sur-apprentissage.....	51
Figure II.7 Schéma d'un neurone formel avec les signaux d'entrée et sortie.....	52
Figure II.8 Représentation des réseaux de neurones mono-couche et multicouches (exemple d'un réseau MLP).....	53
Figure II.9 Exemple d'une classification par k-plus proches voisin (KNN).....	56
Figure II.10 principes de marge et hyperplans séparateurs de SVM.....	57
Figure II.11 Schémas des méthodes d'ensembles de base.....	62
Figure IV.1 Région d'étude (Nord-Est de l'Algérie).....	86
Figure IV.2 Précision de la classification pour l'application de différentes variantes SVM multiclassés utilisant le noyau polynomial par rapport au noyau gaussien dans la classification des précipitations.....	101
Figure IV.3 Différentes stratégies utilisées pour la construction de SVM multiclassés, a) OvA-SVM multiclassés, b) OvO-SVM multiclassés, c) SBT-SVM multiclassés, d) DDAG-SVM multiclassés.....	102

Liste des figures

Figure IV.4 : Schéma d'application de l'algorithme AdaBoost sur les différentes variantes de SVM multiclassées, a) AdaBoost sur OvO-SVM, b) AdaBoost sur OvA-SVM, c) AdaBoost sur SBT-SVM, d) AdaBoost sur DDAG-SVM.....	104
Figure IV.5 : Résultats de classification de la scène de précipitations instantanées (29 novembre 2010 à 12h00 pm), 1ère ligne : classée par radar (à droite), observée dans le canal infrarouge10.8 (à gauche). Entre la 2ème ligne et la 5ème ligne : résultats de classification sans AdaBoost (à gauche), résultats de classification avec AdaBoost (à droite).....	107
Figure IV.6 Précision de la classification en fonction du nombre d'itérations de AdaBoost.....	108
Figure IV.7 Résultats de la classification de la scène instantanée des précipitations(29 novembre 2010 à 12h00 pm), a) classée par AdaOvA-SVM, b) classée par SART, c) classée par RFT, d) classée par CS-RADT, e) classée par ECST.....	115
Figure IV.8 : Résultats de l'estimation des taux de précipitations utilisant AdaOvA-SVM par rapport au pluviomètre, a) échelle journalière, b) échelle mensuelle, c) échelle saisonnière.....	118
Figure IV.9 : Estimation des précipitations en utilisant les différentes variantes (AdaOvA-SVM, SART, RFT, CS-RADT et ECST) comparée aux mesures pluviométriques.....	119

Liste des tableaux

Tableau I.1 Chronologie et propriétés des satellites METEOSAT.....	18
Tableau I.2 Caractéristiques spectrales et propriétés des canaux du radiomètre SEVIRI.....	22
Tableau IV.1 Relation entre l'intensité des précipitations et la réflectivité radar (Z-R).....	88
Table IV.2 Paramètres spectraux choisis comme entrée de notre modèle et leurs valeurs possible et les informations qu'ils fournissent.....	91
Tableau IV.3 Classes des précipitations et intervalles correspondants en dBZ.....	92
Tableau IV.4 :Périodes des bases de données d'apprentissage et de validation avec le nombre de scènes de précipitations instantanées (IPS) et de pixels par classe pour chaque période.....	105
Tableau IV.5. Précision de classification après 0, 100, 200, 300 et 400 itérations de AdaBoost pour les différentes variantes.....	109
Tableau IV.6 Vue d'ensemble des combinaisons dans un tableau de contingence.....	110
Tableau IV.7 Paramètres d'évaluation de la classification des classes convectives et stratiformes pour différentes variantes de SVM multiclassées avec et sans AdaBoost.....	111
Tableau IV.8 Résultats statistiques d'évaluation pour les différentes techniques.....	114
Table IV.9 : Paramètres d'évaluation des différentes méthodes dans le cas des estimations de précipitations.....	120

Introduction

1. Introduction

La mesure des taux de précipitations d'une façon fiable à haute résolution spatio-temporelle est d'une très grande importance pour la compréhension du cycle hydrologique. Elle permet aussi de fournir des éléments de réponse aux problèmes de sécheresse et aux préventions des événements de précipitations intenses. La gestion des ressources en eau et la planification des activités agricoles passent nécessairement par la connaissance des champs de précipitations dans le temps et dans l'espace. La mesure continue et complète des précipitations n'est pas une opération simple, car le manque d'instruments de mesure au sol, tels que les pluviomètres ou les radars météorologiques rend la couverture discontinue et partielle. En effet, ces instruments traditionnellement utilisés pour la mesure directe des précipitations pour leur performance sont limités en termes de couverture spatiale. Des zones difficiles d'accès, telles que les zones désertiques, océanique ou zones montagneuses restent non suffisamment couvertes, voire non couvertes. A cet effet, pour mesurer les précipitations, les chercheurs ont recours aux données issues des satellites météorologiques. Outre leur haute résolution spatio-temporelle permettant une meilleure couverture des régions et leur disponibilité, de grandes quantités d'informations sur les propriétés des nuages sont fournies par les capteurs embarqués sur des satellites météorologiques géostationnaires ou autres instruments.

Toutefois, les données satellites ne sont pas des mesures directes des taux de précipitations. Ainsi, pour tenter d'établir une relation entre les informations satellitaire et les taux de précipitation, plusieurs méthodes et techniques mathématiques ont été développées (Fiedas et al 2008 ; thies et al 2008; Roebelling et al 2009 ; Lazri et al 2013a,b ;2014 ; Lazri et al, 2020 ; Sehad et Aneur, 2020 ; Ouallouche et al, 2018). Dans les premiers travaux d'estimation des précipitations, des techniques statistiques sont employées pour établir une relation entre la température du sommet des nuages (CTT) extraite des canaux infrarouges des données satellites météorologiques et les quantités de précipitations (e.g.Arkin et Childs;1985 ; Negri et Alder 1993; Adler et Negri 1988). Cependant, ces techniques sont adaptées principalement à la détection des précipitations de types convectifs. Elles ne peuvent pas être appliquées à la situation complexe des précipitations frontales (convectives et stratiformes) des régions extratropicales (Levizzani 2003;Thies et al 2008;Feidas et al 2008...). Dans ce contexte, bénéficiant du nouveau satellite MSG (Météosat Seconde

Introduction

Génération) offrant des grandes quantités de données à des résolutions spatio-temporelles et spectrales supérieures aux précédents MFG (Météosat première génération), des méthodes développées incorporent des canaux multi spectraux supplémentaires fournissant des informations sur les propriétés optiques et microphysique des nuages, telles que la phase thermodynamique, épaisseur optique, tailles des gouttelettes des nuages (Nauss et al, 2005 ; Nauss et Kokhanovsky 2006 ; Rebellin et al 2006). En se basant sur le modèle conceptuel selon lequel les nuages stratiformes précipitants sont caractérisés par une grande épaisseur optique et un rayon de particules de nuage suffisamment grand, ces méthodes ont considérablement amélioré les résultats de classifications et d'estimation des intensités de précipitations dans les régions à moyenne latitudes (Lensky et Rosenfeld 2003; Thies et al 2008a,b ; Feidas et Giannakos2011 ; Lazri et al 2013).

Pour tirer profit de la quantité et de la variabilité de ces données satellitaires, des algorithmes de classification basés sur l'apprentissage automatique, tels que les réseaux de neurones artificiels, les machines à vecteurs de support, les forêts aléatoires sont massivement utilisés (Lazri et al, 2020 ; Lazri et Ameer, 2018 ; Sehad et Ameer, 2020). Ils se sont avérés très efficace dans la classification et l'estimation des intensités de précipitations. Toutefois, malgré les résultats obtenus et les niveaux de précision atteints par l'utilisation de ces différents classifieurs d'apprentissage automatique, il y a encore de nombreuses améliorations à apporter. Pour tenter d'optimiser les résultats, certains travaux de recherche ont augmenté les bases de données d'apprentissage (Shtanko et Kulik, 2021). Mais parfois il existe des problèmes de sur-apprentissage où le classifieur ne généralise pas parfaitement une fois appris sur ces ensembles de données d'apprentissage.

De plus, le Deep Learning basé sur les réseaux de neurones profonds présente une option prometteuse où il a montré son efficacité dans la reconnaissance et la détection d'objets sur des images (Kavitha et al, 2022).Le deep learning comme CNN (convolution neural network) opère sur un ensemble de pixels représentant un objet permettant de le détecter et de le reconnaître. L'image ou une partie de l'image subit une répétition de convolution et de pooling (regroupement) avant d'être connectée à un MLP (perceptron multicouche) qui génère la réponse (objet détecté ou reconnu) en sortie. Dans le cas d'une classification des précipitations qui se fait à l'échelle du pixel, l'apprentissage profond de type CNN ne peut pas être appliqué. Quant au DNN (deep neural network), le problème de cette option est la complexité et le volume de données nécessaires à son fonctionnement alors que, dans le cadre

Introduction

de cette étude, la quantité de données fournies par le MSG est insuffisante pour la mise en œuvre du DNN.

D'autres travaux effectuent des ajustements au niveau des architectures des classifieurs ou des modifications au niveau des paramètres d'entrées sans pour autant obtenir un véritable apport (Gupta et Raza, 2019).

La tendance aujourd'hui est orientée vers la recherche de techniques ensemblistes pour augmenter la précision de la classification. Dans cette optique, l'introduction de méthodes d'ensemble a énormément amélioré les classifieurs. Ces méthodes combinent plusieurs classifieurs appelés 'apprenants faibles' pour classer de manière itérative l'ensemble des données, puis donner une réponse fiable. Les méthodes d'ensemble les plus connus sont les algorithmes de Boosting (Schapire, 1990 ; Schapire et Singer, 1999). L'idée générale des algorithmes de Boosting est de combiner séquentiellement une succession de prédicteurs apprenants, où chaque prédicteur suivant tente de corriger les erreurs de son prédécesseur. Parmi ces algorithmes, on cite la méthode Adaptive Boosting (AdaBoost) de (Freund et Schapire, 1997) qui est une version de l'algorithme Boosting. Le grand succès d'AdaBoost peut être attribué à sa capacité à maximiser la marge sur un ensemble d'entraînement, ce qui conduit à améliorer les performances du classifieur.

Pour la mise au point d'AdaBoost, le choix de classifieur faible reste controversé. Un classifieur faible doit juste faire mieux que le hasard. Dans le cas des SVM, les résultats des travaux de recherche divergent. Certains travaux soutiennent l'idée d'utiliser le SVM comme un classifieur binaire faible (e.g. Cheng et Jhan, 2013). D'autres considèrent que le SVM est un classifieur fort ; l'implémentation de Boosting sur ce classifieur introduira des désaccords dans les décisions entre les différents modèles appris (e.g. Wickramaratna et al, 2001). L'utilisation d'un classifieur qui est déjà fort (comme SVM) comme apprenant de base dans AdaBoost ne semble pas offrir d'avantages significatifs en termes de généralisation. Mais une version affaiblie de SVM peut toujours être utile comme classifieur de base.

En effet, dans cette étude, nous considérons que les performances de AdaBoost ne dépendent pas uniquement du classifieur mais également de l'environnement dans lequel le classifieur évolue (par exemple : la variante du classifieur, le type de données : données déséquilibrées, présence de valeurs aberrantes, etc.)

Ainsi, nous visons deux objectifs, d'une part, nous testons l'implémentation de l'algorithme AdaBoost sur SVM multiclassés. À cette fin, nous avons considéré quatre stratégies de SVM

Introduction

multiclasses, à savoir un-contre-tous SVM (OvA-SVM), un-contre-un-SVM (OvO -SVM), Slant Binary Tree SVM (SBT-SVM) et Decision Directed Acyclic Graph SVM (DDAG-SVM) qui doivent être appliquées avec l'algorithme AdaBoost. D'autre part, nous effectuons des classifications et des estimations des précipitations en choisissant les meilleurs paramètres d'entrée multispectraux du satellite MSG.

Les modèles sont appliqués aux données de MSG en exploitant les propriétés optiques et microphysiques des nuages comme paramètres d'entrée des modèles développés. Les données du radar météo de Sétif sont utilisées pour l'apprentissage et la validation de la technique.

Pour bien illustrer le travail que nous avons réalisé, nous avons organisé notre thèse en quatre chapitres comme suit :

Le premier chapitre est consacré à la présentation des différents types de précipitations se produisant dans notre zone d'études (Nord de l'Algérie). Des instruments utilisés pour la collecte des données météorologiques, à savoir les instruments de mesure au sol et de télédétection font aussi l'objet de ce chapitre.

Dans le deuxième chapitre, nous avons défini le concept mathématique de l'apprentissage automatique ainsi quelques méthodes de classification basées sur ce concept. Nous avons aussi décrit des méthodes ensemblistes utilisés pour améliorer certains de ces algorithmes de classification.

Dans le troisième chapitre, nous avons passé en revue quelques modèles appliqués à l'estimation des précipitations. Nous avons classé ces modèles selon le type de données utilisées dans le traitement et le concept de fonctionnement, à savoir statistique ou basé sur l'apprentissage automatique.

Dans le dernier chapitre, nous présentons la méthodologie que nous avons élaborée pour les estimations des précipitations. Elle est basée sur l'implémentation de l'algorithme AdaBoost pour optimiser les résultats de classification obtenus en appliquant le SVM multiclasse avec la stratégie un contre tous.

Nous terminons notre thèse par une conclusion en présentant quelques perspectives pour ce travail.

CHAPITRE I

Instruments de mesure et d'observation des précipitations

I.1. Préambule :

Selon l'Organisation météorologique mondiale (OMM), les précipitations sont les produits solides ou liquides résultant de la condensation de la vapeur d'eau qui tombent des nuages, elles comprennent plusieurs types comme la pluie, la grêle, la neige etc.

La mesure des taux de ces précipitations est importante dans de nombreux domaines, tels que la météorologie, l'économie et le secteur agricole, elle est aussi essentielle pour l'étude des variations climatiques et de la protection de l'environnement. Toutefois, la configuration géographique et la topographie de chaque région, ainsi que les mouvements des nuages précipitants et leurs activités, rendent la mesure et la quantification des précipitations plus complexes. Ces paramètres augmentent considérablement la variabilité spatio-temporelle des précipitations et donc la difficulté de leurs estimations.

En région méditerranéenne, le climat est influencé à la fois, par le climat subtropical et par celui des systèmes de moyennes latitudes (Trigo et al. 2006; Alpert et al. 2006). Il en résulte que le climat de cette région est parmi les plus instables et d'une pluviométrie extrêmement variable. En effet, les systèmes convectifs et les systèmes stratiformes peuvent coexister dans les mêmes cellules pluvieuses. Des telles situations rendent les prévisions et les estimations des plus complexes. Malgré la fiabilité des mesures en utilisant les moyens au sol, leurs couvertures spatiales sont insuffisantes.

L'objectif de ce chapitre est de présenter les types de précipitations dans la région méditerranéenne. Nous décrivons aussi les instruments de mesure directe et indirecte, à savoir les pluviomètres, les radars météorologiques et satellites météorologiques employés pour l'observation et la collecte des informations sur les précipitations.

I.2. Précipitations et leurs types :

Dans les travaux et études visant à mesurer ou estimer les taux de précipitations par moyens de télédétection (satellites et radar), la classification des nuages précipitants et leurs séparation en différents types, constitue une étape primordiale pour une meilleure estimation.

Il existe principalement deux types de précipitations en fonction de la nature des mouvements d'air verticaux. Les précipitations convectives qui sont caractérisées par une vitesse verticale (ascendante) d'air supérieure ou égale à la vitesse de chute des gouttelettes d'eau ou de glace, tandis que pour les précipitations stratiformes la vitesse verticale de l'air est inférieure à la vitesse de chute des gouttelettes d'eau ou de glace.

I.2.1. Précipitations convectives

Elles sont causées par les vitesses ascendantes de l'air humide qui sont généralement plus grandes que celles des chutes des particules, ce qui permet d'avoir une condensation de la vapeur d'eau générant des nuages à grand contenu en eau. Les processus de givrage et coalescence favorisent le développement des précipitations liquides et solides, qui tombent au sol lorsque ces particules précipitantes deviennent suffisamment lourdes sous l'effet de la pesanteur. Les précipitations résultantes par ce processus sont caractérisées par leur fort développement vertical et elles sont localisées, de courte durée et de forte intensité (Houze,1997).

I.2.2. Précipitations stratiformes

Elles sont causées par un soulèvement lent et à grande échelle de l'humidité, due à la vitesse verticale ascendante de l'air humide qui est inférieure à la vitesse de chute des particules d'eau, ce qui permet la formation des nuages à contenu en eau plus faible que ceux des précipitations convectives. Les précipitations stratiformes se caractérisent par des sommets de nuages moins verticaux et relativement chauds, avec une extension horizontale importante et homogène, dont les intensités varient de faibles à modérées, selon que leurs origines soit les front chauds ou froids, et des durées plus longues que celles des précipitations convectives (Waymire et al,1984).

I.3. Instruments de mesure et d'observation des précipitations

On peut classer les instruments de mesures ou d'observations météorologiques en deux catégories : instruments de mesure directe, qui consiste à recueillir les précipitations en un

endroit donné et à mesurer la quantité recueillie pendant un certain intervalle de temps, et instruments de mesures indirectes (observation) par télédétection.

Nous allons décrire les principaux appareils pouvant servir à la mesure et l'observation des précipitations, à savoir le pluviomètre, le pluviographe pour les mesures directes ainsi que le radar et les satellites météorologiques pour les mesures indirectes par télédétection.

I.3.1. Instruments de mesure au sol

Pour la mesure au sol des taux de précipitations, ça consiste en la mesure directe des quantités d'eau tombée durant une certaine période de temps sur une zone bien déterminée, couvertes par ces instruments de mesure au sol qui recueillent les précipitations d'une manière directe, et expriment généralement ces taux de précipitations en lame d'eau précipitée par unité de surface horizontale ($1 \text{ mm} = 1 \text{ l} / \text{m}^2 = 1 \cdot 10^{-3} \text{ m}^3 / \text{m}^2$).

On peut distinguer deux types de ces instruments, en l'occurrence le pluviomètre et le pluviographe. Le premier est un dispositif de mesure simple, constitué d'une surface horizontale formant le récipient collecteur et qui nous donne la quantité de pluie précipitée durant l'intervalle de temps séparant deux relevés. Le pluviographe, comme pour le pluviomètre, collecte l'eau précipitante par l'intermédiaire d'un dispositif collecteur, mais relié à un système permettant l'enregistrement du volume d'eau collectée en fonction du temps sur un support numérique. Pour plus de détails sur les types d'instruments de mesure des précipitations au sol voir les références (Seck,2020 ; Lazri,2013).

Les mesures fournies par ce type d'instruments sont des mesures directes et fiables. Cependant, ces mesures présentent des inconvénients dues à la couverture spatiale insuffisante, et leur représentativité est limitée au maximum à '1 km' à cause la grande variabilité spatiale des précipitations. Ces difficultés sont accentuées, notamment dans les zones difficiles d'accès telles que les zones montagneuses, au dessus des océans et dans les régions désertiques.

A cet effet, les chercheurs se sont tournés vers les données des radars et satellites météorologiques avec leurs larges couvertures spatiales et leur bonne résolution spatio-temporelle, permettant ainsi d'améliorer considérablement la connaissance de la distribution des précipitations.

I.3.2. Instruments de mesures à distance

En utilisant des techniques d'acquisition d'informations à distance basées sur les propriétés du rayonnement électromagnétique, la télédétection spatiale permet d'observer le système terre, océans et atmosphère. Les capteurs satellitaires et d'autres type comme le radar météorologique, permettent de mesurer l'énergie réfléchie et (ou) émise par la surface de la terre, les océans ou par l'atmosphère, ce qui permet donc d'analyser et de suivre l'évolution de ces milieux au fil du temps.

La télédétection est particulièrement importante pour la mesure et la classification de précipitations dans de vastes régions océaniques et régions isolées, où la couverture en mesure au sol est rare ou insuffisante. Les instruments de télédétection comprennent deux catégories à savoir les instruments de télédétection active et de télédétection passive.

La télédétection active consiste à transmettre une impulsion de rayonnement électromagnétique, et à mesurer le rayonnement réfléchi par la cible des mesures ou de l'observation, dans ce cas le radar météorologique est un exemple de capteur actif. Quant à la télédétection passive, elle mesure le rayonnement électromagnétique venant des surfaces surveillées et étudiées, reçus par les imageurs et capteurs à micro-ondes et infrarouges, embarqués à bord des satellites.

I.3.2.1. Radars météorologiques :

Le terme 'RADAR' est l'acronyme de l'expression anglaise « Radio Detection And Ranging » (détection par ondes radio et mesures de distance). Le radar émet des ondes électromagnétiques se propageant à la vitesse de la lumière dans l'atmosphère. Lorsque celles-ci rencontrent un obstacle, elles sont en partie réfléchies et l'antenne radar capte un écho en retour pour l'analyser et obtenir des informations sur la cible source de rétrodiffusion. Le radar a été introduit en aéronautique durant la seconde guerre mondiale, dans le but de suivre et de détection des objets volants. Depuis cette époque, les systèmes radar se sont sophistiqués, et les radars météorologiques sont apparus pour occuper une place essentielle dans le dispositif d'observation de l'atmosphère et des études climatologiques.

Les premiers radars météorologiques ont été réalisés entre 1950 et 1980 pour scruter les nuages et les précipitations. Ils émettent des ondes électromagnétiques à fortes fréquences,

puis reçoivent les échos rétrodiffusés par les cibles contenus dans ces nuages (pluies, grêles, neiges).

L'analyse des échos reçus, et le traitement de ces ondes permet de localiser les nuages précipitants et les précipitations, définir leurs types et identifier ceux qui pourraient causer des dégâts climatiques considérables (Wolff et Vaillant ,2011).

Ainsi le radar météorologique constitue un outil indispensable dans les organisations et centres météorologiques mondiaux, et dans les projets de recherches scientifiques liés à l'observation de l'atmosphère, ainsi qu'à l'élaboration des prévisions météorologiques et climatologiques.

L'utilisation quantitative des données radar météorologique, comme données d'entrée des modèles de prévision hydrologique, ou pour l'apprentissage et la validation des méthodes d'estimation des précipitations utilisant les données satellite, est devenue de plus en plus répandu dans les récentes études. Ceci est grâce à sa couverture spatiale et temporelle continue des zones où il ya manque d'instrument de mesure au sol.

I.3.2.1.1. Mesures par le radar météorologique

Pour mesurer les intensités de précipitations, les radars sont dotés d'unité de calcul et traitement, il s'agit de l'unité au niveau de laquelle s'effectuent les différentes opérations de traitement des signaux, rétrodiffusés reçues par l'antenne du radar. Pour plus de détails et information sur les principes et équations de fonctionnement du radar météorologique, le lecteur pourra se référer aux travaux de (Lazri,2013 ; Seck,2020) et (Annexe B) .

Ces traitement des signaux effectués par le radar permettent d'obtenir de l'information utile sur les cibles étudiées (nuages précipitants), tel que le facteur de réflectivité (Z) à partir du quel on pourra obtenir les taux de précipitations et les natures des nuages. A partir de l'intensité du signal rétrodiffusé reçu, le système radar pourra déterminer une estimation sur la taille ou la forme des objets diffuseurs, qui sont les nuages précipitants et donc avoir de l'information sur l'intensité des précipitations. La localisation de ces nuages est obtenue selon l'orientation de l'antenne et la mesure du temps écoulé entre l'émission du signal et la réception de la fraction rétrodiffusée.

Le radar calcule alors la distance qui le sépare de ces précipitations, cette distance ' r ' qui est donnée par la relation :

$$r = \frac{c\tau}{2} \quad (1.1)$$

Ou τ est le temps mis par les ondes rétrodiffusées, pour faire leurs allers retours entre le radar et la cible.

Pour déterminer l'intensité, un facteur de réflectivité appelé Z est utilisé permettant de traduire la capacité de la cible qui est le nuage précipitant à renvoyer l'énergie électromagnétique qu'elle reçoit par le radar lors de l'émission des impulsions.

En supposant que la distribution des gouttelettes d'eau dans les nuages précipitants soit continue. En appliquant l'approximation de Rayleigh, la réflectivité des gouttes de pluie exprimant l'intensité de réflexion, qui est proportionnelle aux taux de précipitations dans le volume sondé, est donnée par (Sauvageot, 1994) :

$$\eta = \frac{1}{V} \cdot \frac{\pi^5}{\lambda^4} \cdot |K|^2 \cdot \int_{D_{min}}^{D_{max}} D^6 N(D) dD \quad (1.2)$$

C'est l'équation de Rayleigh pour la réflectivité des gouttes de pluie, avec : $D \ll \lambda$. Dans cette expression $N(D) dD$ exprime le nombre de particules de dimension comprises entre D et $D + dD$, V est le volume d'impulsion (volume sondé) contenant diverses gouttes de pluie de diamètre D et $|K|^2$: est la réflectivité du milieu considéré.

L'intégrale qui intervient dans l'équation (1.2) est, par définition, le facteur de réflectivité Z caractérisant le champ des précipitations (Sauvageot 1994).

$$Z = \int_{D_{min}}^{D_{max}} D^6 N(D) dD \quad (1.3)$$

La réflectivité (Z) exprimé usuellement en (mm^6/m^3).

Dans le cas de spectre discret, le facteur de réflectivité est donné par la relation suivante (Sauvageot 1992) :

$$Z = \sum_i \Delta Z_i = \sum_i [N_i(D_i) \Delta D_i] D_i^6 = \sum_i n_i(D_i) D_i^6 \quad (1.4)$$

Le facteur Z est une caractérisation de la répartition moyenne des diffuseurs dans le volume V , il est donné par la relation.

$$Z = \frac{1}{V} \sum_i D_i^6 \quad (1.5)$$

En pratique, on exprime η en (cm^{-1}), λ en cm, et Z en (mm^6 / m^3). En fait, traduire le facteur Z en unités logarithmiques définies par comparaison avec un niveau de référence de ($1 mm^6 / m^3$) est plus pratique, sa valeur en dBZ est alors donnée par (Sauvageot, 1992) :

$$Z(\text{dBZ}) = 10 \log [Z (mm^6 / m^3)] \quad (1.6)$$

Ainsi, il est possible de déterminer l'intensité de pluie à partir du facteur de réflectivité. En effet, l'intensité de pluie notée R , qui est exprimée en quantité (volume) de pluie sur unité de surface durant une unité (période) de temps, est donnée par la relation suivante (Sauvageot, 1992) :

$$R = \frac{\pi}{6} \int_{D_{min}}^{D_{max}} D^3 V_i(D) N(D) dD \quad (1.7)$$

Où $V_i(D)$ est la vitesse de chute d'une goutte de pluie de diamètre D en l'absence de courant d'air ascendant vertical. Elle est en fonction du diamètre de cette goutte de pluie et peut être approchée par la relation suivante (Sauvageot 1992):

$$V_i(D) = K \cdot D^\alpha \quad (1.8)$$

Où K et α sont des coefficients d'ajustement.

Pour déduire les champs d'intensité de pluie à partir des données et réflectivité radar météorologique, il y a une relation d'étalonnage permettant de convertir la réflectivité radar (Z) en intensité de pluie R , cette relation est de la forme suivante :

$$Z = a R^b \quad (1.9)$$

Où Z est la réflectivité du radar météorologique, et R est le taux des précipitations exprimé en (mm/h).

Les constantes 'a' et 'b' sont déterminées expérimentalement, elles sont essentiellement dépendantes des facteurs et propriétés du climat de la région étudiée, ainsi que des caractéristiques des précipitations et des nuages dans cette région.

Divers modèles ont été développés pour déterminer ces constantes 'a' et 'b', et établir la relation entre la réflectivité radar et les taux de précipitations à partir de travaux expérimentaux, nous citons :

- Modèle Marshall-Palmer (1948) pour les précipitations stratiformes : $Z = 200 \cdot R^{1.6}$.
- Modèle de Jones(1956) pour les précipitations convectives : $Z = 360 \cdot R^{1.35}$.

I.3.2.2. Satellites météorologiques :

Les satellites météorologiques sont des plates-formes en orbite, qui disposent à leurs bords des instruments (capteurs et imageurs) pouvant détecter le rayonnement et l'énergie thermique de l'atmosphère et des surfaces observées. Les satellites météorologiques constituent un système perpétuel d'observation de l'atmosphère terrestre. En plus de renvoyer des images vers la terre, les satellites météorologiques peuvent nous permettre de déterminer la température et le contenu de vapeur d'eau à différentes altitudes dans l'atmosphère. Ils peuvent aussi contribuer à la surveillance de la couche d'ozone.

Les observations satellitaires constituent un apport primordial aux différents systèmes de prévisions météorologiques numériques, surtout dans les zones où la couverture des radars météorologiques et pluviomètres est manquante ou inexistante, ces observations permettent également aux scientifiques et prévisionnistes de détecter le développement de phénomènes météorologiques susceptibles de causer des dégâts, et donc mieux protéger l'environnement. La capacité des satellites météorologiques à fournir des données et mesures de l'espace, pour le soutien aux études de l'évolution climatique prend de plus en plus d'importance.

Les capteurs à bord de ces satellites réagissent à deux types d'énergie rayonnante de base :

- La lumière visible produite par le soleil qui est réfléchiée par les surfaces observées de la terre et les nuages et qu'est retournée au satellite ;
- l'énergie infrarouge ou thermique provenant de surfaces observées terrestres ou aquatiques de la Terre et des sommets de nuages.

Pour coordonner la dynamique d'évolution rapide de lancement de multiples séries de satellites météorologiques de différentes nations et organisations spatiales, un système complet fut mis en place par les membres du Groupe de Coordination des Satellites Météorologiques (GCSM). Ce réseau opérationnel de satellites météorologiques est sous la responsabilité de quatre opérateurs satellitaires : EUMETSAT (Europe), JAXA (Japon), FKA (Russie) et NASA (Etats-Unis).

Finalement, le système mondial de satellites météorologiques opérationnels est composé d'une constellation d'au moins cinq satellites, régulièrement espacés autour de l'équateur en orbite géostationnaire, et au moins deux satellites en orbite quasi-polaire pour bénéficier de la complémentarité et les avantages de ces deux types. (Figure I.1).

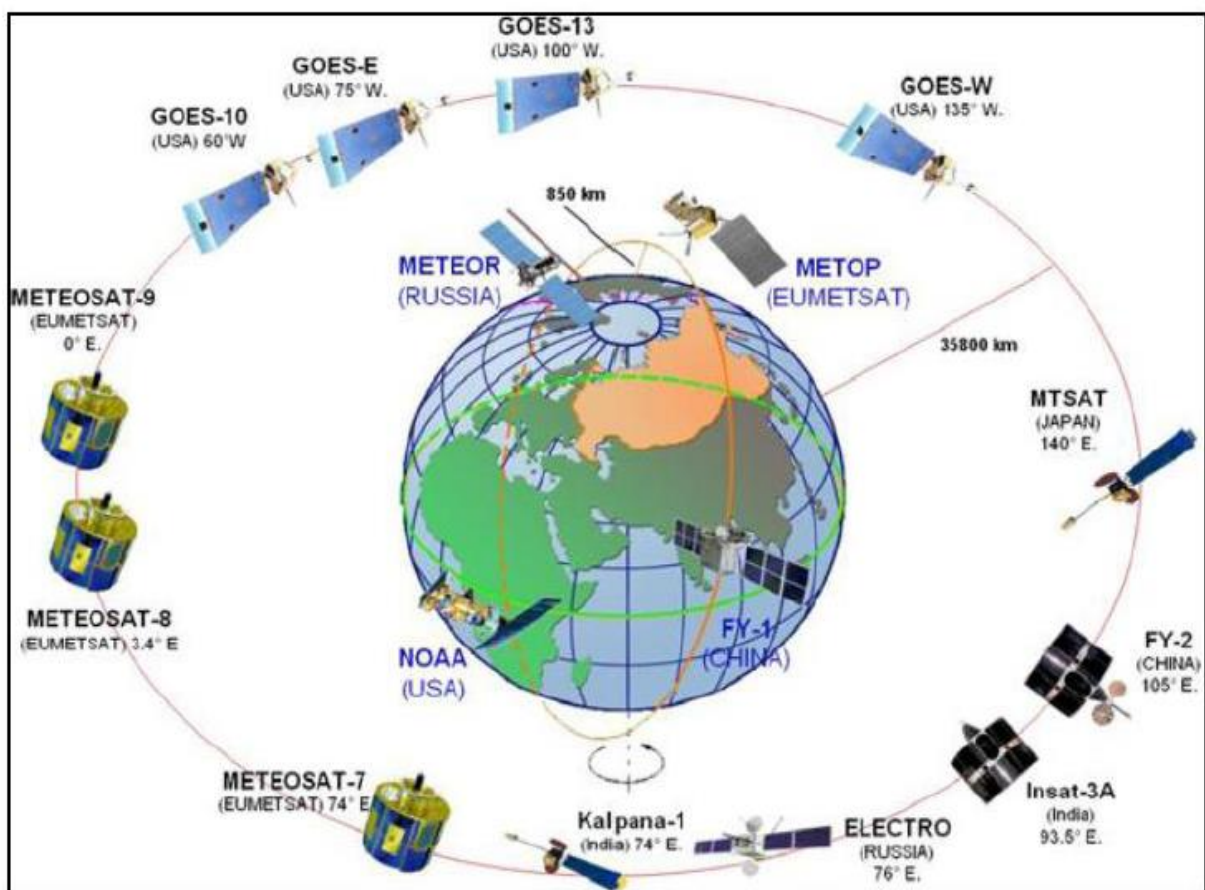


Figure I.1 : Système mondiale des satellites d'observation météorologiques et leurs distributions (Source Eumetsat).

I.3.2.2.1. Satellites météorologiques à orbite polaires :

Ce sont des satellites à défilement, qui tournent autour de la terre sur une orbite beaucoup plus basse, et dont le plan passe pratiquement près des pôles de la Terre, dans des trajets Nord et Sud et légèrement incliné (par exemple : NOAA pour les américains et METOP pour les européens par EUMETSAT) voir figure I.1.

Ils observent la Terre à seulement 850 km d'altitude. Ce qui leurs permet de fournir des informations avec une meilleure résolution spatiale que les satellites géostationnaires de l'ordre de 2 km, vu qu'ils sont plus proches. Et leurs orbites leurs permettent surtout l'observation des régions qui ne sont pas couvertes par les satellites géostationnaires qui sont les zones polaires.

Les satellites en orbite polaire peuvent balayer une bande de la Terre, en prenant moins de deux heures pour compléter une orbite. À chaque passage, ils examinent une bande d'une largeur d'environ 1900 kilomètres qui se situe plus à l'Ouest à cause de la rotation de la Terre vers l'Est, les imageurs de ces satellites permettent de surveiller les masses nuageuses sur l'ensemble de la terre, mais surtout dans les régions polaires (nord de la Scandinavie par exemple) qui ne sont pas couvertes par les satellites géostationnaires.

Comme ils sont plus rapprochés de la surface, on peut même en extraire des informations sur le vent selon la forme et le déplacement des nuages. Mais, vu qu'ils ne couvrent pas continuellement la même surface terrestre, ils ont un usage plus limité pour surveiller la météo en temps réel.

En effet ils passent uniquement deux fois par jour au-dessus du même endroit de moyenne ou basse latitude, ce qui ne permet pas l'utilisation de l'animation des images pour suivre le mouvement et l'évolution des masses nuageuses d'une manière efficace.

Les satellites à orbite polaires portent une plus grande variété d'instruments que les satellites géostationnaires, et peuvent fournir des images de la planète avec plus de précision, mais moins fréquemment. Cette complémentarité des propriétés des deux types de satellites météorologiques, explique leur utilisation combinée dans le système mondial des satellites météorologiques.

La figure I.2 montre le plan de l'orbite d'un satellite à défilement.

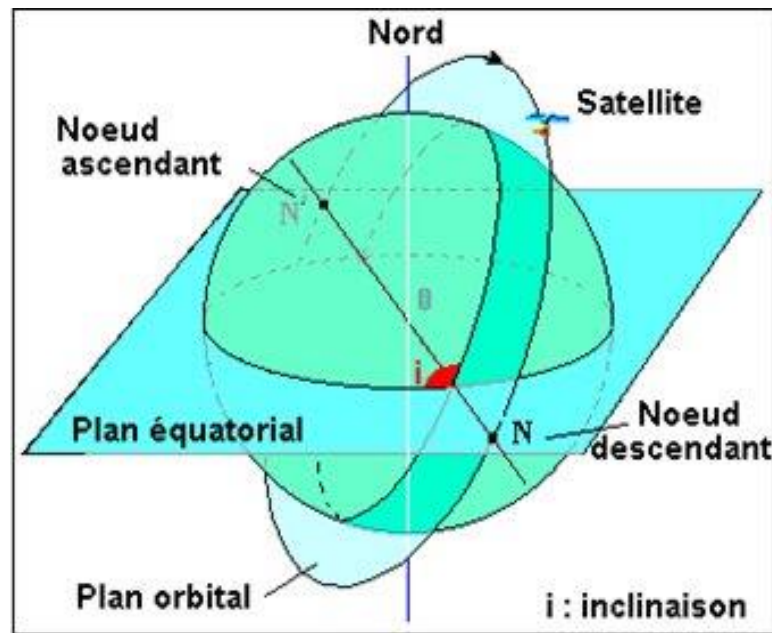


Figure I.2 : Orbite d'un satellite à défilement.

I.3.2.2.2. Satellites météorologiques géostationnaires :

Les satellites géostationnaires peuvent couvrir la même zone terrestre à tout moment car ils tournent dans le même sens que la Terre et à la même vitesse. Ils sont situés à une altitude de 36 000 km au-dessus de l'Équateur sur des endroits spécifiques à la même latitude que l'équateur 0° . Leur altitude élevée et leur position fixe par rapport à la terre leur permettent de fournir des images couvrant une grande (large) surface de la même zone de la terre à un rythme élevé, une caractéristique adéquate pour le suivi de l'évolution des nuages et des phénomènes météorologiques.

La résolution des satellites géostationnaires est maximale à leur sous-point, le point de l'équateur à la verticale duquel ils sont situés. Cette résolution diminue en allant vers les bords du disque terrestre, à cause de la parallaxe de l'angle de visée de plus en plus rasant. Ce type de satellites fournit aux météorologistes des données et images de la terre avec une bonne résolution temporelle.

La (figure I.3) montre les zones de la terre couvertes par chacun des satellites géostationnaires.

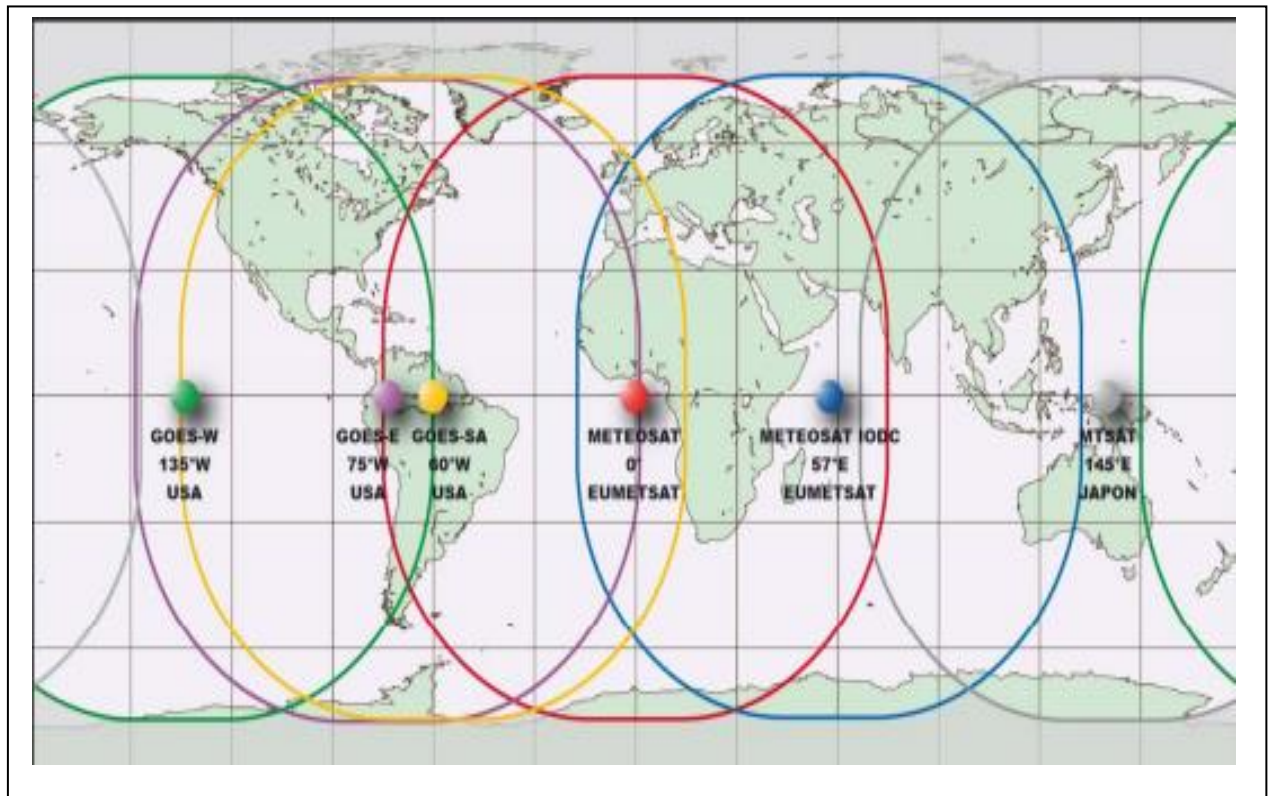


Figure I.3 : Zones couvertes par les différents satellites météorologiques géostationnaires (source EUMETSAT [Eumet]).

Comme le montre la figure I.3 les régions d'Europe et Afrique du nord y compris notre zone d'études sont couvertes par les satellites Meteosat, que nous décrivons dans la partie suivante.

I.3.2.2.3. Satellites Météorologiques Météosat :

Le programme Météosat fait partie du système mondial d'observation de l'atmosphère, mis en place par l'Organisation Météorologique Mondiale OMM au milieu des années 1970, qu'est un ensemble de satellites météorologiques en orbite géostationnaire, qui se positionnent à différentes longitudes et permet d'assurer une couverture global de la planète voir (figure I.1).

Les satellites METEOSAT sont des satellites météorologiques européens, dont le propriétaire et l'opérateur sont l'organisation européenne chargée de l'exploitation des systèmes opérationnels de satellites météorologiques EUMETSAT et l'agence spatiale européenne (ESA : European Space Agency). L'utilisation principale de leurs images est destinée à la prévision météorologique. Cependant, d'autres utilisations sont aussi envisageables, tel que l'observation de couverture et surface agricoles ou surveillance de l'environnement etc. Ils sont appelés satellites géostationnaires car ils sont immobiles au-dessus d'un point particulier de la terre, du fait que leur vitesse de rotation angulaire est identique à celle du globe terrestre.

Les satellites METEOSAT sont situés à environ 36000 km d'altitude au-dessus de l'équateur. Leur champ de vision couvre l'Afrique, l'Europe et les extrémités de l'Asie et de l'Amérique méridionale (Figure I.4). Les images acquises par ces satellites couvrent toujours la même zone du globe terrestre.

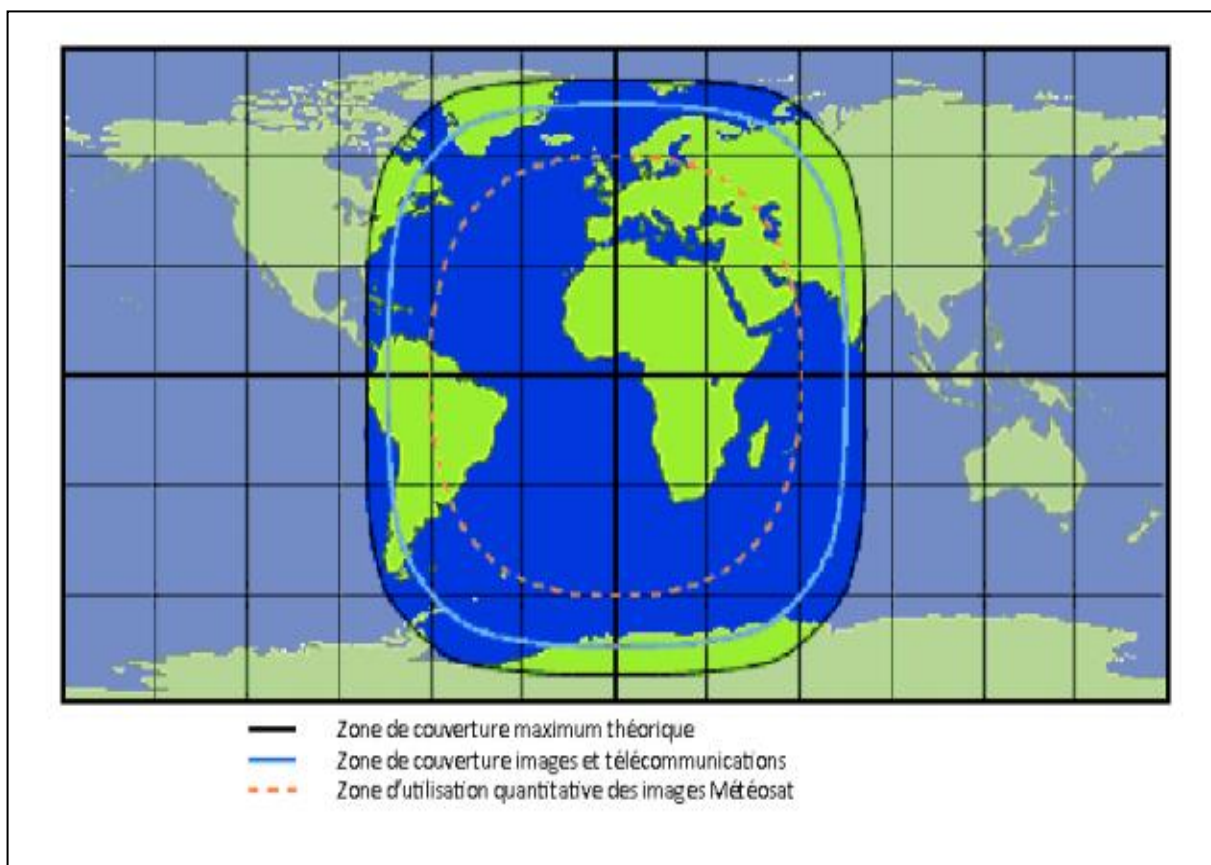


Figure I.4 : zone de couverture des satellites METEOSAT.

Le premier Météosat a été lancé en 1977, suivi d'une série de plusieurs satellites, dont le dernier des satellites de seconde génération MSG avant le lancement de la troisième génération de ces satellites (Meteosat Third Generation MTG), est le satellite de seconde génération « MSG4 » a été lancé en 2015.

Le tableau I.1 donne l'année de lancement, ainsi que les caractéristiques principales de chacun des satellites METEOSAT depuis le début de leur lancement.

satellite	Date de lancement	Nombre de canaux	Fréquence d'acquisition	Résolution spatiale
METEOSAT1	Novembre 1977	3	30 min	5km (2.5 pour HRV)
METEOSAT 2	Juin 1981	3	30 min	5km (2.5 pour HRV)
METEOSAT 3	Juin 1988	3	30 min	5km (2.5 pour HRV)
METEOSAT 4	Mars 1989	3	30 min	5km (2.5 pour HRV)
METEOSAT 5	Mars 1991	3	30 min	5km (2.5 pour HRV)
METEOSAT 6	Novembre 1993	3	30 min	5km (2.5 pour HRV)
METEOSAT 7	Septembre 1997	3	30 min	5km (2.5 pour HRV)
MSG1	Aout 2002	12	15 min	3km (1 pour HRV)
MSG2	Décembre 2005	12	15 min	3km (1 pour HRV)
MSG3	Juillet 2012	12	15 min	3km (1 pour HRV)
MSG4	Juillet 2015	12	15 min	3km (1 pour HRV)

Tableau I.1 Chronologie et propriétés des satellites METEOSAT.

I.3.2.2.4. Satellite Météosat deuxième génération (MSG Meteosat Second Generation)

Le premier exemplaire de la seconde génération de satellites européens de veille météorologique (Meteosat Second Generation MSG-1) a été mis en orbite le 28 août 2002 par un lanceur européen Ariane 5 qui a décollé du Centre Spatial de Kourou, en Guyane française, près de vingt-cinq ans après le lancement du tout premier satellite Météosat, en novembre 1977. Ce satellite devenu opérationnel le 9 janvier 2004 prenant le nom de Météosat-8, puis c'était au tour de MSG-2 d'être lancé le 22 décembre 2005, avant d'être déclaré en service en juillet 2006 sous le nom de Meteosat-9. Quant aux deux derniers satellites du programme Météosat Seconde Génération, ils ont été lancés le 5 juillet 2012 pour MSG-3 (Meteosat10) et le 15 juillet 2015 pour MSG-4 (Meteosat11).

Les satellites MSG sont deux fois et demie plus gros que leurs prédécesseurs. Ce sont des satellites cylindriques de 3,22 mètres de diamètre pour 2,4 mètres de haut (Figure I.6). Ils sont équipés de nouveaux capteurs plus puissants et plus précis que ceux qu'étaient à bord de la génération précédente de satellites Meteosat, pour l'observation en continu de l'atmosphère

terrestre avec de meilleures résolutions, fournissant jusqu'à l'horizon 2025 une multitude de données indispensables à la compréhension et à la modélisation des activités climatiques de notre planète.

Les satellites MSG sont stabilisés par rotation de 100 tours/minute autour d'un axe parallèle à l'axe nord-sud de la terre, leurs radiomètre SEVIRI visent la terre par l'intermédiaire d'un miroir, et balayent une bande étroite de la surface de la terre à chaque révolution du satellite voir (figure I.5).

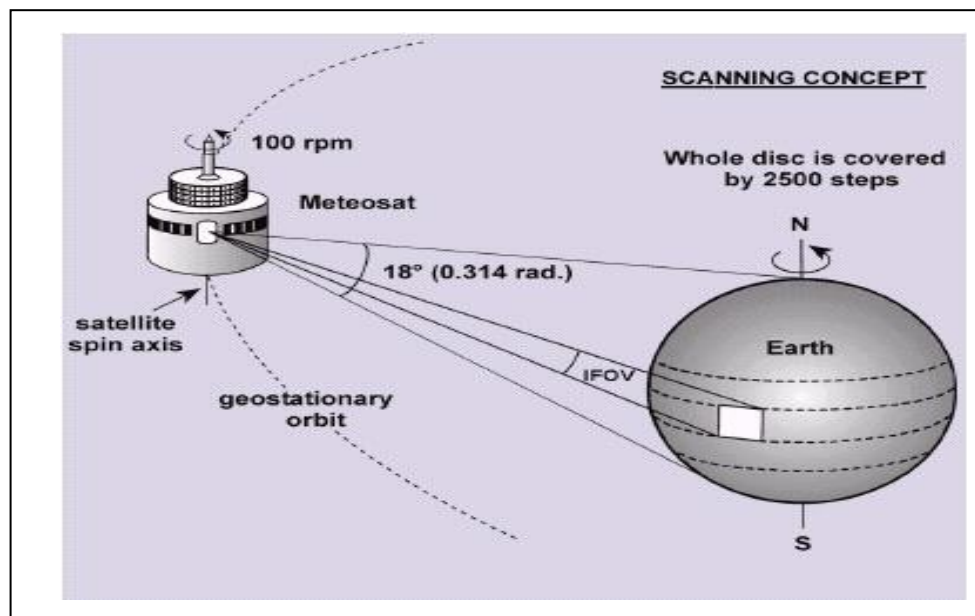


Figure I.5 : Concept de scan du satellite METEOSAT (source EUMETSAT [Eumet]).

I.3.2.2.4.1. Radiomètre SEVIRI (Spinning Enhanced Visible and Infra Red Imager)

Le radiomètre SEVIRI du satellite MSG, permet l'acquisition d'images toutes les 15 minutes, au lieu de 30 minutes pour METEOSAT de première génération, ce qui constitue une grande amélioration.

Il comporte onze canaux, qui exploitent des bandes spectrales allant du visible à l'infrarouge thermique, avec une résolution spatiale améliorée de 3km, et un canal panchromatique (HRV) à haute résolution spatiale de (1km) comparé à seulement trois canaux pour le radiomètre du Meteosat première génération MFG (Figure I.6).

Cet enrichissement du spectre des observations de SEVIRI représente une avancée significative dans l'amélioration des modèles météorologiques numériques qui utilisent ces données. De plus, le MSG permet aux climatologues et aux météorologues de repérer plus facilement le déclenchement des phénomènes à évolution rapide comme les orages, les tempêtes de neige ou les bancs de brouillard, en réduisant la fréquence de rafraîchissement des données de 30 à 15 minutes.

La résolution spatiale des canaux infrarouge passe de 5km à 3km dans le MSG, et aussi une résolution au sol dans le spectre visible pour le canal (HRV) de 1 km au lieu de 2,5 km précédemment, ce qui permet l'observation et le suivi de phénomènes localisés (Eumetsat,2005).

En plus du radiomètre SEVIRI, les satellites MSG embarquent à leurs bords le radiomètre GERB qui fournit des données essentielles sur le bilan radiatif de la Terre, représenté par le rapport entre la quantité d'énergie reçue du Soleil par notre planète et celle rayonnée en retour.

Le rapport signal-bruit qui était élevé avec celui de METEOSAT première génération, a été amélioré sur MSG grâce à l'élargissement des bandes spectrales.

La figure I.6 représente la comparaison entre MSG et MFG (Meteosat First Generation).

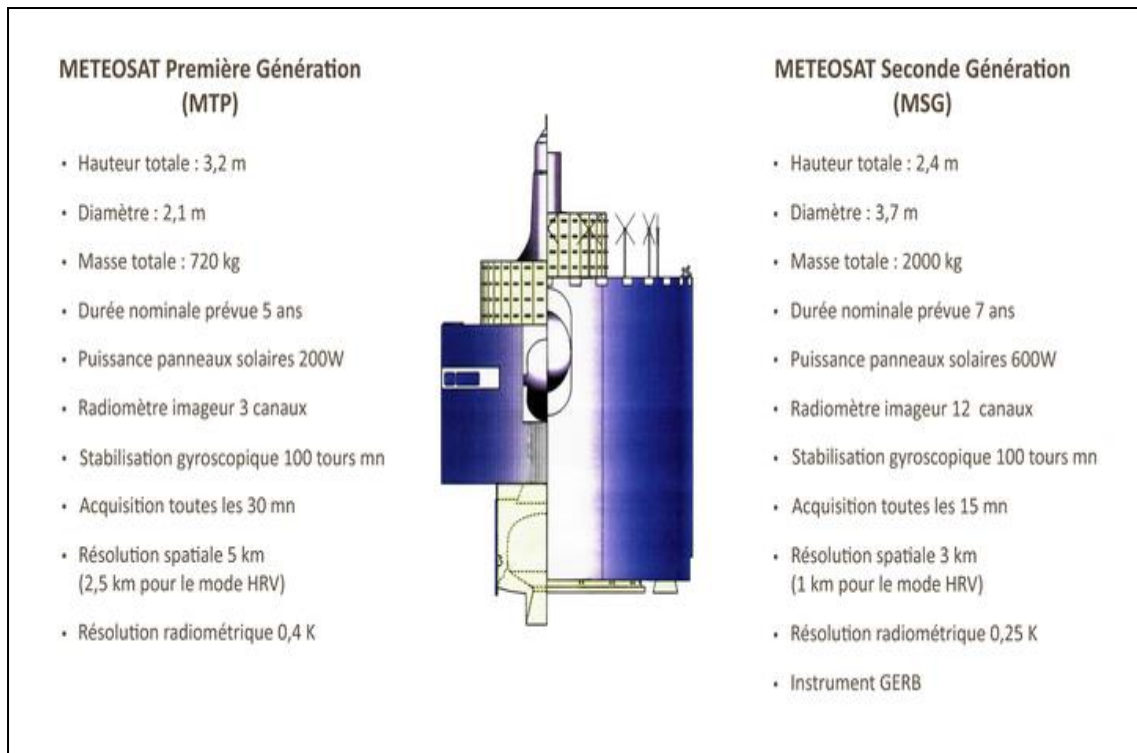


Figure I.6 : comparaison entre les satellites MFG et MSG de METEOSAT (Eumetsat2002).

Le tableau I.3 donne les noms des douze canaux du radiomètre SEVIRI de MSG, la bande spectrale et la longueur d'onde centrale de chaque canal et sa fonction (Eumetsat, 2006).

Numéro du canal	Nom du canal	Bandes spectrales(longueurs d'onde en μm)	Fonctions des différents canaux
C1	VIS 0.6	0.56 à 0.71	Détection et suivi des nuages. Surveillance de la surface des terres et des aérosols. Leur combinaison permet d'obtenir des indices de végétation.
C2	VIS 0.8	0.74 à 0.88	
C3	NIR1.6	1.5 à 1.78	Distinction entre la neige et les nuages, entre les nuages de glace et les nuages d'eau. Informations sur les aérosols.
C4	IR 3.9	3.48 à 4.36	Détection nocturne et propriétés des nuages bas et du brouillard. Mesures nocturnes des températures du sol et de la mer. Détection des feux de forêt.
C5	WV 6.2	5.35 à 7.15	Mesure de la vapeur d'eau à mi-atmosphère. Chaque canal correspond à une couche atmosphérique différente. Calculs des vents.
C6	WV 7.3	6.85 à 7.85	
C7	IR 8.7	8.30 à 9.10	Informations quantitatives sur les cirrus. Distinction entre les nuages de glace et les nuages d'eau.
C8	IR 9.7	9.38 à 9.94	Surveillance de l'ozone total. Mesure de sa concentration dans la basse stratosphère. Indication des champs de vents à cette altitude.
C9	IR 10.8	9.80 à 11.80	Infrarouge thermique. Mesure de la température des nuages et de la surface. Calcul des vents. Estimation de l'instabilité atmosphérique.
C10	IR 12.0	11.00 à 13.00	
C11	IR 13.4	12.40 à 14.40	Absorption du CO ₂ . Estimation de l'instabilité atmosphérique. Estimation de la température de la basse troposphère. Mesure de la hauteur des nuages semi-transparents.
C12	HRV	0,50 à 0,90	Large bande dans le visible comme le canal VIS de Météosat mais une résolution spatiale plus fine : 1km au lieu de 2,5 km.

Tableau I.2 Caractéristiques spectrales et propriétés des canaux du radiomètre SEVIRI.

Le MSG produit deux sortes de fichiers images : image 1.0 et image 1.5.

Les données d'image de niveau 1.0 sont acquises par le radiomètre SEVIRI des satellites MSG et sont transmises au niveau de la station de Darmstadt en Allemagne. Les données de niveau 1.5 sont dérivées des données de niveau 1.0, où EUMETSAT corrige en temps réel chaque image de niveau 1.0 reçue, de tous les effets radiométriques et géométriques, et la géolocalise à l'aide d'une projection standardisée (Lacaze et Bergès, 2005 ; Eumetsat 2004).

Les différents canaux d'observation du radiomètre SEVIRI de MSG sont décrits dans les paragraphes qui suivent (Eumetsat, 2006).

i. Canaux de SEVIRI

Le radiomètre SEVIRI de MSG fournit des images de la partie qu'il couvre du disque terrestre, dans 12 canaux différents (figure I.7).

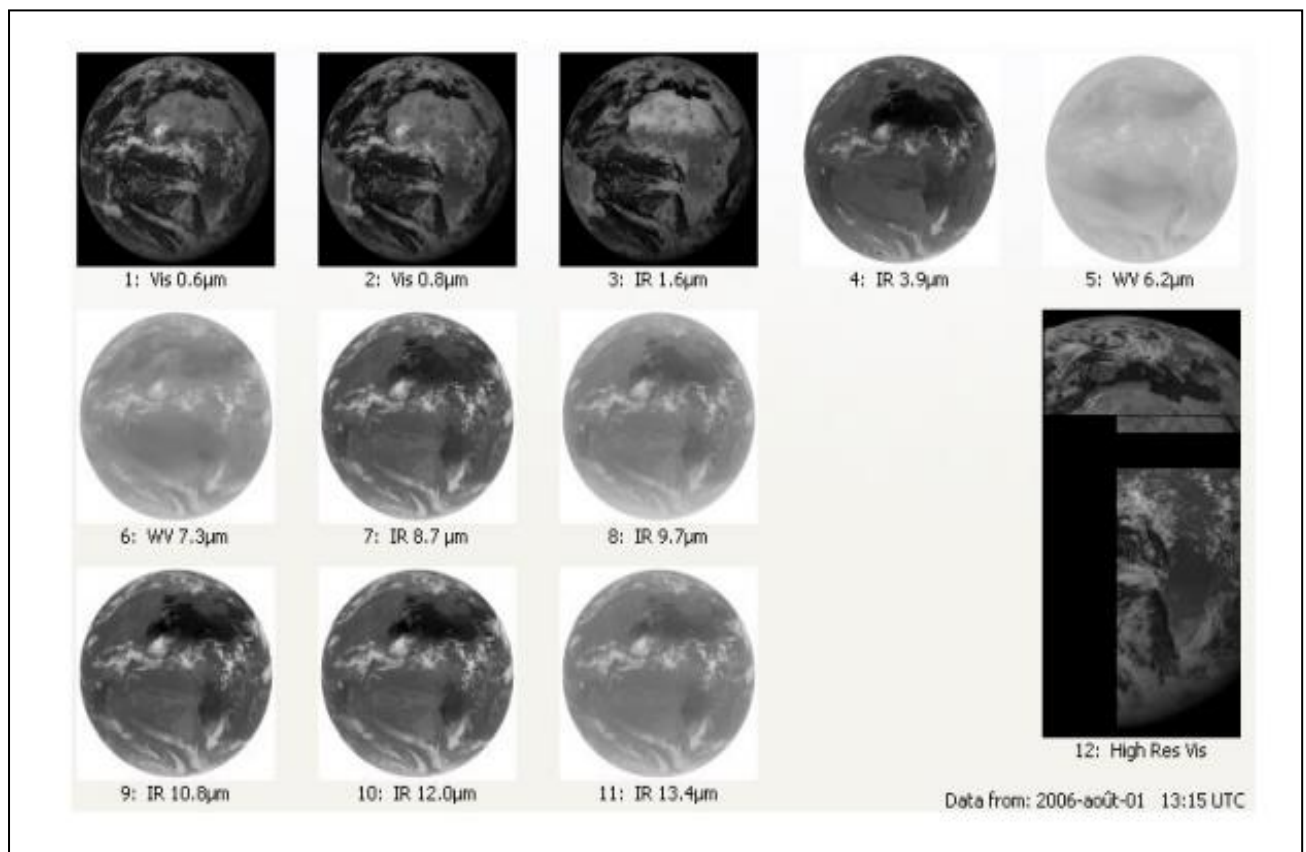


Figure I.7 : images acquises à partir des différents canaux de MSG-SEVIRI le 01 aout 2006 à 13 :00 UTC (source Eumetsat)

✓ Les canaux visibles :

MSG1 possède deux canaux visibles : un dans la longueur d'onde $0.6 \mu\text{m}$ et l'autre dans la longueur d'onde $0.8 \mu\text{m}$. Ces deux canaux sont perceptibles par l'œil humain. Les images provenant de ces canaux, sont des images monochromes composées de niveaux de gris, et qui ont des caractéristiques similaires à l'image du visible de Météosat première génération.

Les images visibles représentent la quantité de lumière visible rétro diffusée par les nuages ou la surface de la terre ; dans ces images la neige et les nuages apparaissent en blanc et les zones sans nuages apparaissent en noir. Les nuages épais sont plus brillants que les nuages fins. Pendant la nuit, les images visibles sont complètement noires et ne peuvent donc pas être utilisées.

La taille des images est de 3712×3712 pixel, l'utilisation des images de ce canal visible permet la détection et le suivi des masses nuageuses durant le jour, ainsi que l'étude des terres émergées, les aérosols, et la surveillance de la végétation.

✓ Le canal High Resolution Visible (HRV):

Les images visibles de haute résolution sont générées par le satellite MSG, elles sont acquises dans la bande spectrale $0.4 - 1.1 \mu\text{m}$. La résolution spatiale de ces images est (1 km^2) alors que pour les autres images composées par le capteur elle est de (3 km^2), mais représentant une zone réduite du globe terrestre

Ce type d'images est bien adapté pour l'étude des nuages d'un point de vue structurale, les détails texturaux sont mieux distingués sur de telles images, cela permet la détection de structures nuageuses de petite taille. La taille de ces images HRV est de 11136×5568 pixels, où chaque pixel représente une réflectance entre 0 et 100%.

✓ Les canaux vapeur d'eau (Water Vapor WV) :

MSG1 possède deux canaux vapeur d'eau : WV $6.2 \mu\text{m}$ et WV $7.3 \mu\text{m}$, ces canaux maintiennent le canal vapeur d'eau de Météosat première génération.

Les images «vapeur d'eau», représentent une mesure du rayonnement infrarouge influencée par la vapeur d'eau dans l'atmosphère, cela aide à distinguer les zones sèches et les zones humides. Les canaux vapeur d'eau, permettent de mesurer la vapeur d'eau dans la

mésosphère, et d'estimer l'altitude des nuages semi-transparents. Les deux canaux séparés permettent d'obtenir des informations sur le contenu en vapeur d'eau à différents niveaux de la troposphère. Un niveau de gris clair entraîne une température d'émission faible, donc une humidité élevée qu'a absorbée le rayonnement, et un niveau de gris sombre entraîne une température d'émission élevée, donc une humidité faible.

✓ **Le canal proche Infra Rouge NIR1.6 :**

Ce canal est situé dans la partie dite proche infrarouge du spectre électromagnétique. Il aide à différencier entre les nuages et la neige, et entre nuages de glace et nuages d'eau. Il fournit également des informations sur la présence d'aérosols dans l'atmosphère.

✓ **Le canal IR3.9**

Il se trouve dans une zone du spectre électromagnétique où la lumière du soleil influe sur le rayonnement reçu. Par conséquent, il s'agit d'un canal d'émission thermique et il est également partiellement sensible à la réflexion du rayonnement solaire. De ce fait, il est nécessaire de prendre quelques précautions lorsqu'on interprète les images et données de ce canal. En effet, durant le jour, cette bande spectrale est dominée par l'énergie solaire réfléchie qui lui donne l'apparence d'une image dans la partie visible du spectre. En revanche, pendant la nuit, l'énergie ne provient que du rayonnement thermique de la Terre, qui présente des caractéristiques dans la partie infrarouge du spectre.

Ce canal est principalement utilisé pour détecter les nuages bas, les brouillards nocturnes et les incendies forestiers. De plus, il peut être utilisé pour mesurer les températures de surface du sol et de la mer.

✓ **Le canal IR8.7 :**

Les images dans ce canal apportent essentiellement des informations sur les masses nuageuses minces des cirrus et permet de faire la différence entre les nuages de glace et d'eau.

✓ **Le canal IR9.7 :**

Ce canal est influencé par la concentration d'ozone dans la partie basse de la stratosphère. Il sert à mesurer l'ozone total et à évaluer sa variabilité diurne. Il est susceptible de suivre les tendances de l'ozone en tant qu'indicateur des champs de vent à ce niveau.

✓ **Les canaux IR10.8 et IR12.0 :**

Ces canaux sont situés dans l'infrarouge thermique, ils réagissent aux rayonnements émis par les sommets des nuages et la surface terrestre donnant l'information sur leur température. Ce qui nous permet l'estimation de la hauteur des sommets des nuages et leur développement vertical. Ils permettent également le suivi des nuages pour déterminer les vents atmosphériques et estimer l'instabilité atmosphérique.

✓ **Le canal IR13.4 :**

Ce canal se trouve dans la zone du spectre où les molécules de dioxyde de carbone (CO₂) absorbent le rayonnement. Il est utilisé pour évaluer l'instabilité atmosphérique et apporte des informations sur la température de la basse troposphère.

ii. Principe d'acquisition d'images par SEVIRI

Le SEVIRI est un radiomètre imageur à balayage, employé par le satellite MSG afin d'obtenir des images dans une plage spectrale plus étendue, grâce à la rotation du satellite autour de son axe d'inertie principal. Le satellite tourne à une vitesse de 100 tours par minute autour d'un axe parallèle à l'axe Nord-Sud de la Terre, ce qui fait qu'un tour complet du satellite sur son axe dure 0,6 seconde.

Le radiomètre de Météosat utilise un miroir pour viser la terre, et à chaque révolution du satellite, il balaie une étroite bande de la surface terrestre (voir figure I.8). L'angle de balayage ou d'ouverture correspondant de 18° est décrit en 30 ms. Au cours des 570 ms suivantes, le télescope cible l'espace et cette période est utilisée pour réajuster les miroirs afin de les préparer pour le prochain cycle de scan. Ainsi, au tour suivant il balaie au sol une bande adjacente à la précédente, mais plus au nord (Eumetsat, 2004).

Il faut environ 1250 révolutions du satellite pour obtenir toutes les 3712 lignes d'image, à l'aide de trois détecteurs, réalisées en 12 minutes et 30 secondes. Les 2 minutes et 30 secondes suivantes sont consacrées au retour du miroir à sa position initiale. La phase de non-acquisition de données est utilisée pour calibrer les détecteurs, ce qui permet d'obtenir un cycle de répétition d'imagerie de 15 minutes, donnant ainsi un total de 96 acquisitions d'images par jour.

Le satellite maintient sa stabilité sur son orbite grâce à sa propre rotation autour d'un axe parallèle à celui de la terre, à une vitesse de 100 tours par minute. On désigne le pixel comme la région élémentaire observée à chaque instant. Il s'agit d'une zone au sol d'environ 1km^2 pour le canal HRV et de 3km^2 pour les canaux IR et WV, ainsi que pour les trois canaux visibles restants, au point sub-satellite pour les satellites de seconde génération (MSG). Le radiomètre mesure l'énergie radiative dans diverses bandes spectrales pour chaque pixel. Une fois numérisées, ces mesures sont envoyées à une station au sol pour être traitée, puis envoyée à la communauté des utilisateurs.

Les données ainsi acquises constituent les images METEOSAT. Dans le cas de MSG, la zone d'observation ou l'image complète est constituée de $11\ 136 \times 5568$ pixels pour le canal HRV, tandis que pour les canaux IR, WV et les autres canaux visibles les images sont composés de $3\ 712 \times 3\ 712$ pixels.

Le principe d'acquisition d'une image de la terre par le radiomètre SEVIRI est illustré par la figure I.8.

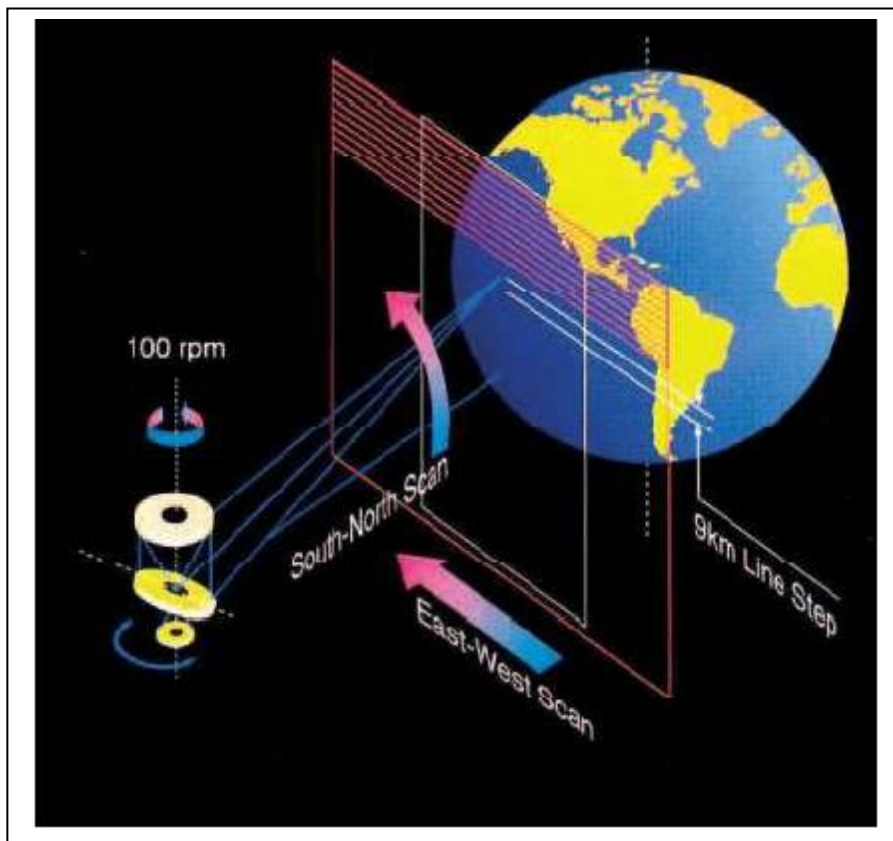


Figure I.8 : Principe d'acquisition des images des satellites Météosat.

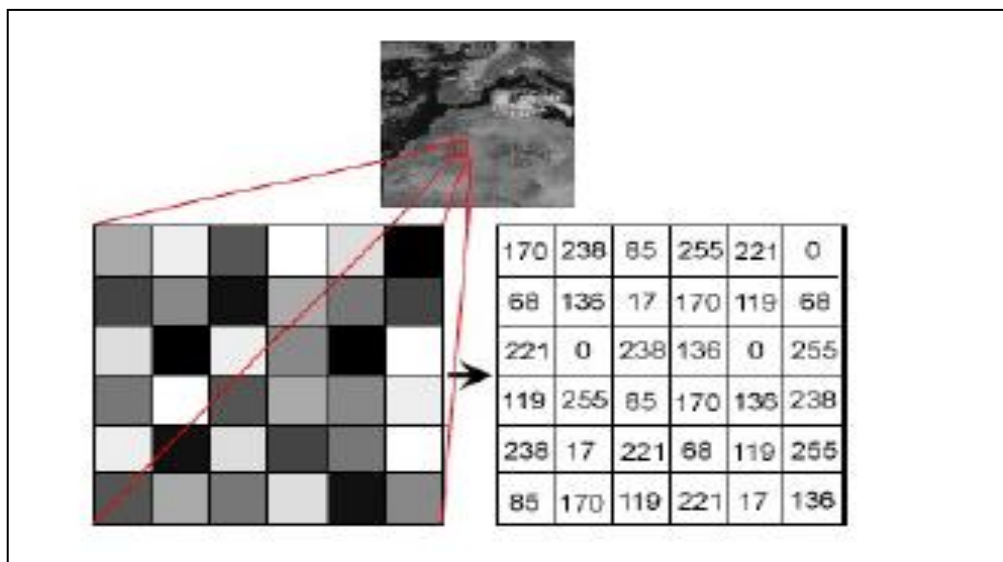


Figure I.10 : La résolution radiométrique.

I.4.Discussion

Dans ce chapitre, nous avons fait une présentation des différents types de précipitations. La différenciation entre ces types est très utile dans l'estimation des précipitations à partir des données de télédétection. Comme nous l'avons bien indiqué, les précipitations de type convectif sont intenses et de courte durée. En revanche, les précipitations stratiformes sont modérées et durent plus longtemps.

Nous avons aussi décrit dans ce chapitre les instruments de mesures et de télédétection des précipitations, tels que les outils au sol comme les pluviographes et pluviomètres, les radars météorologiques qui utilisent le principe de télédétection active et les satellites météorologiques couvrant de vastes zones fonctionnant avec le principe de télédétection passive. Le satellite Meteosat Second Generation (MSG) qui est le satellite géostationnaire utilisé pour la collecte des informations sur les nuages précipitants couvrant notre zone d'étude (Nord de l'Algérie) a été présenté, ainsi que les différents canaux actifs et le type d'information qu'ils fournissent.

Cependant, les informations issues de ce type de satellite ne sont pas directement liées aux taux de précipitations. A cet effet, pour estimer les précipitations, des méthodes sont élaborées permettant de relier les propriétés optiques et microphysiques des nuages aux taux de précipitations. Particulièrement, ces dernières années, en raison de leur efficacité, les modèles mathématiques basés sur l'apprentissage automatique sont largement utilisés à cette fin. Dans le prochain chapitre, nous introduisons les concepts de base de l'apprentissage automatique.

Chapitre II

Etat de l'art sur l'apprentissage automatique (Machine Learning)

II.1. Préambule :

Ces dernières années, dans la plupart des domaines qui traitent les images satellitaires et de télédétection, les méthodes basées sur l'apprentissage automatique (Machine Learning) sont largement utilisées (Maxwell et al 2018). Vu le caractère aléatoire, complexe et multidimensionnel de ces données, les méthodes mathématiques classiques ne peuvent plus les traiter efficacement. Pour optimiser donc les traitements de données, les chercheurs ont opté pour l'application des méthodes basées sur l'apprentissage automatique. Il s'agit d'apprendre à un modèle mathématique dans un premier temps et puis il sera appliqué une fois apprenant. Par conséquent, la compréhension des concepts et des outils utilisés en apprentissage automatique (Machine Learning ML) pour résoudre ces types de problèmes fera l'objet de ce chapitre.

En effet, en premier lieu, nous abordons ici le concept d'apprentissage automatique et les domaines pour lesquels il est utilisé, ses différents types d'algorithmes et comment modéliser un problème avec les méthodes Machine Learning. Dans un second lieu, nous donnons une liste non exhaustive de méthodes d'apprentissage automatique. Nous présentons aussi les méthodes ensemblistes utilisées pour optimiser le machine learning. Le deep learning qui se présente comme une option moderniste dans le domaine de l'apprentissage automatique sera aussi décrit.

II.2. Définition de l'apprentissage automatique (Machine Learning):

Le Machine Learning ou apprentissage statistique est un champ d'étude de l'intelligence artificielle, qui se fonde sur des approches mathématiques et statistiques pour donner aux machines la capacité d'apprendre à partir de données. Il est devenu incontournable qu'il s'agisse de transformation numérique des entreprises, de Big Data ou des différentes applications de traitement de données aujourd'hui. L'expression apprentissage automatique (ou Machine Learning en anglais), a été créée par Arthur Samuel en 1959 (Samuel, 1959).

L'apprentissage automatique permet de résoudre des problèmes sans avoir été explicitement programmés pour.

Cet apprentissage automatique est généralement utilisé, lorsque la programmation explicite d'un algorithme pour résoudre une tâche ou modéliser un phénomène est trop complexe ou difficile, et la généralisation sur tous les cas rencontrés n'est pas possible par un algorithme spécifique. Ceci est dû essentiellement au nombre trop élevé et divers de règles ou de situations à décrire ainsi qu'à la grande taille de données à traiter. Vu que la disponibilité de grands ensembles et quantité de données issues de différents types de sources et capteurs, est une caractéristique de la science moderne, où l'analyse des données est devenue une composante importante de divers domaines. L'apprentissage automatique permet d'avoir un système qui s'optimise en fonction de l'environnement, des données traitées, les expériences et les résultats observés. Il fait appel à la faculté de généralisation. Le Machine Learning s'applique dans la modélisation des données et des différents liens complexes ou non linéaires entre les entrées et sorties des systèmes étudiés créant ainsi une relation optimisée.

Le Machine Learning est employé dans de nombreux aspects de la technologie moderne, allant de la biotechnologie à l'ingénierie des voitures autonomes et des appareils intelligents aux moteurs de recherche, et de la reconnaissance de caractères et de formes à la recherche en génomique, vers la classification et traitement des images de différents types et résolution et issues de différents capteurs (satellites, ou capteurs d'imagerie médicales etc.), ainsi que dans les domaines de surveillance de la terre par télédétection.

Et plus récemment dans ces dernières décennies, les tentatives de faire progresser l'apprentissage automatique et ses capacités, notamment dans les domaines des réseaux de neurones, ont mené à un sous-domaine de l'apprentissage automatique qui est l'apprentissage profond 'Deep Learning'. La (figure II.1) montre la relation entre l'intelligence artificielle, machine learning et deep learning.

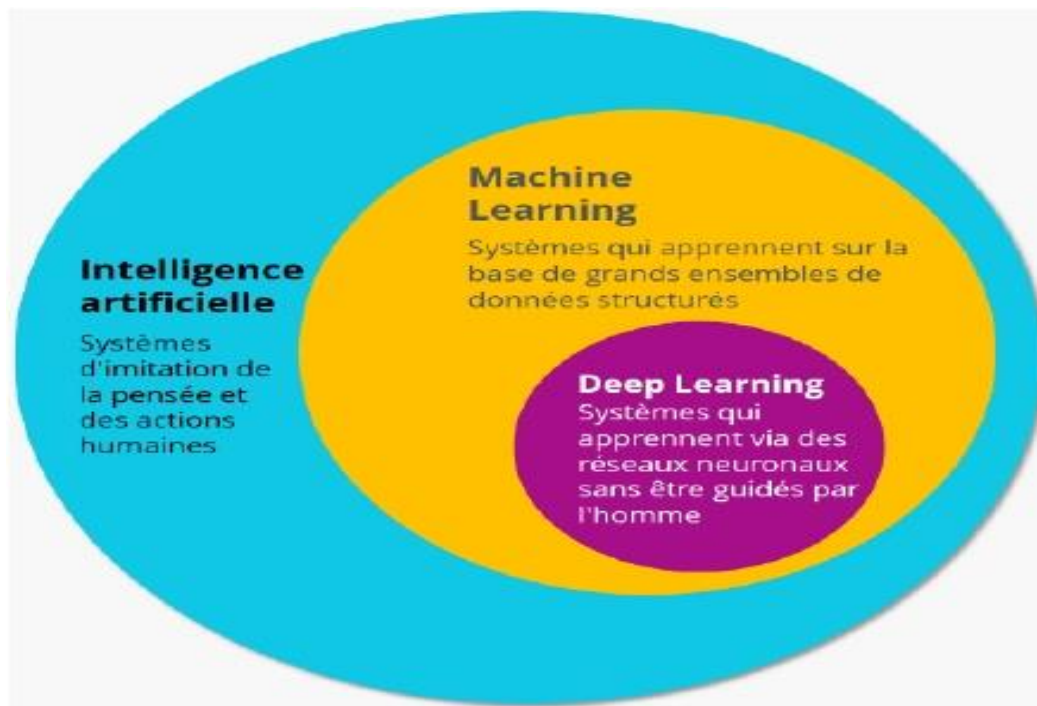


Figure II.1 : la relation entre l'Intelligence Artificielle, le Machine Learning et le Deep Learning.

II.3. Application de l'apprentissage automatique :

L'apprentissage automatique est utilisé dans un vaste champ d'applications afin d'identifier les tâches pour lesquelles les méthodes d'apprentissage automatique peuvent être utiles. Pour un bon fonctionnement de l'apprentissage automatique, Brynjolfsson et Mitchell (2017) proposent les critères suivants :

- La tâche en question implique une fonction qui cartographie des entrées bien définies à des sorties bien définies ;
- De grands ensembles de données existent ou peuvent être créés, contenant des entrées-sorties paires;
- La tâche en question fournit une rétroaction claire, avec des objectifs clairement définissables et métriques.

- La tâche ne nécessite pas d'explications détaillées sur la façon dont la décision a été faite;
- La tâche a une tolérance à l'erreur et n'a pas besoin d'être correcte de manière prouvée ou fournir des solutions optimales.

Ces critères sont des lignes directrices utiles, pour décider si les méthodes d'apprentissage automatique sont adaptées à une tâche ou un domaine donné. Lorsque les données concernant le phénomène étudié sont relativement abondantes, mais que les connaissances en fonctionnement du phénomène sont peu accessibles, ou peu développées, l'apprentissage automatique (machine learning) est utilisé.

II.4. Types d'apprentissage automatique :

Il existe plusieurs types d'apprentissage automatique, qui peuvent être classés selon le mode d'apprentissage qu'ils emploient, et les différentes catégories des problèmes auxquels ils s'intéressent, ainsi que les différents éléments et données disponibles pour la composition de chaque type de ces algorithmes. Donc, nous définissons ici ces différents types d'algorithmes de l'apprentissage automatique :

II.4.1. Apprentissage automatique supervisé :

L'algorithme de ce type d'apprentissage automatique, est entraîné en utilisant une base de données d'apprentissage composée d'échantillons de cas réels traités et validés. Le but de cette étape est de modéliser les liens et déterminer des corrélations entre les données d'entrée des vecteurs ou échantillons à traiter avec leurs données de sorties (variables à prédire) ou étiquettes, pour pouvoir appliquer cette connaissance extraite sur des vecteurs d'entrées avec des sorties inconnues.

Chaque exemple d'apprentissage, appelé aussi instance, est un couple d'entrée-sortie $(x_n, y_n) : 1 \leq n \leq N$.

Avec N : le nombre total des exemples (échantillons). $x_n \in X$ et $y_n \in Y$ et Où :

X : est l'ensemble d'attributs, ou vecteurs d'entrées à traiter (discrets ou continues),

Y : est l'ensemble des valeurs de sortie cible (étiquettes) à prédire, Y peut être discrète ou continue.

En apprentissage supervisé, selon la nature de la variable cible, on distingue deux types de tâches :

- **la classification** : dans le cas où la variable cible (à prédire) est discrète, $Y = \{1, \dots, I\}$.
Ce qui revient à attribuer une classe (ou étiquette) à chaque vecteur x d'entrée.
- **la régression** : quand la variable cible à prédire est continue : $Y \in \mathbb{R}$.
Exemple : prédire le taux de précipitation en (mm/h).

II.4.2 Apprentissage non supervisé :

Dans l'apprentissage non supervisé, les données ne sont pas étiquetées. Pour ce type d'apprentissage, les données d'apprentissage ne contiennent pas de variable cible comme le cas de l'apprentissage supervisé. L'algorithme d'apprentissage doit découvrir par lui-même le lien entre les données, et il doit trouver comment catégoriser ces données. Les algorithmes de regroupement (clustering) sont les types d'algorithmes principaux composants l'apprentissage non-supervisé. Ces algorithmes cherchent à séparer les données d'entrée en un nombre donné de groupes. Chaque élément du groupe doit avoir des caractéristiques proches ou communes de celles des éléments du même groupe.

Parmi les algorithmes d'apprentissage non-supervisé les plus courants, on peut citer l'algorithme des k-moyennes (K-means).

II.4.3. Apprentissage semi-supervisé

Il s'agit d'un hybride entre l'apprentissage supervisé et non supervisé, qui implique l'apprentissage des étiquettes à partir d'un jeu de données partiellement étiqueté. Cette méthode permet d'éviter de devoir étiqueter l'intégralité des exemples d'apprentissage, ce qui réduit le travail de prétraitement des données. Ceci est important lorsque l'étiquetage de ces données nécessite une quantité de travail humain, de puissance de calcul et d'espace de mémoire supplémentaire. Le fait d'inclure un grand nombre de données non étiquetées au cours du processus d'entraînement a tendance à améliorer la performance du modèle final tout en réduisant le temps et les coûts de construction et permet de réduire l'effet des erreurs humaines lors de l'étiquetage.

II.4.4. Apprentissage par renforcement

L'apprentissage par renforcement implique l'interaction avec son environnement. Un système de récompenses favorise les choix positifs du modèle et pénalise les choix négatifs. Le modèle est initialisé aléatoirement et adapte sa stratégie en fonction de son expérience

durant la phase d'entraînement. Pour ce faire, un simple retour des résultats, appelé signal de renforcement, est nécessaire pour enseigner à la machine comment elle doit agir.

Ce modèle de fonctionnement est complexe et peut être modélisé par l'utilisation des réseaux de neurone. D'où l'utilisation d'apprentissage profond (Deep Learning) qu'est aujourd'hui la solution de prédilection pour l'apprentissage par renforcement, en raison de la complexité des modèles. Ce qui distingue l'apprentissage par renforcement des autres types, ce sont l'apprentissage par essais et erreurs et la récompense différée.

Ce type d'apprentissage est souvent utilisé dans le cadre de la théorie des jeux et des véhicules autonomes. La figure II.2 illustre les différents types d'apprentissage automatique

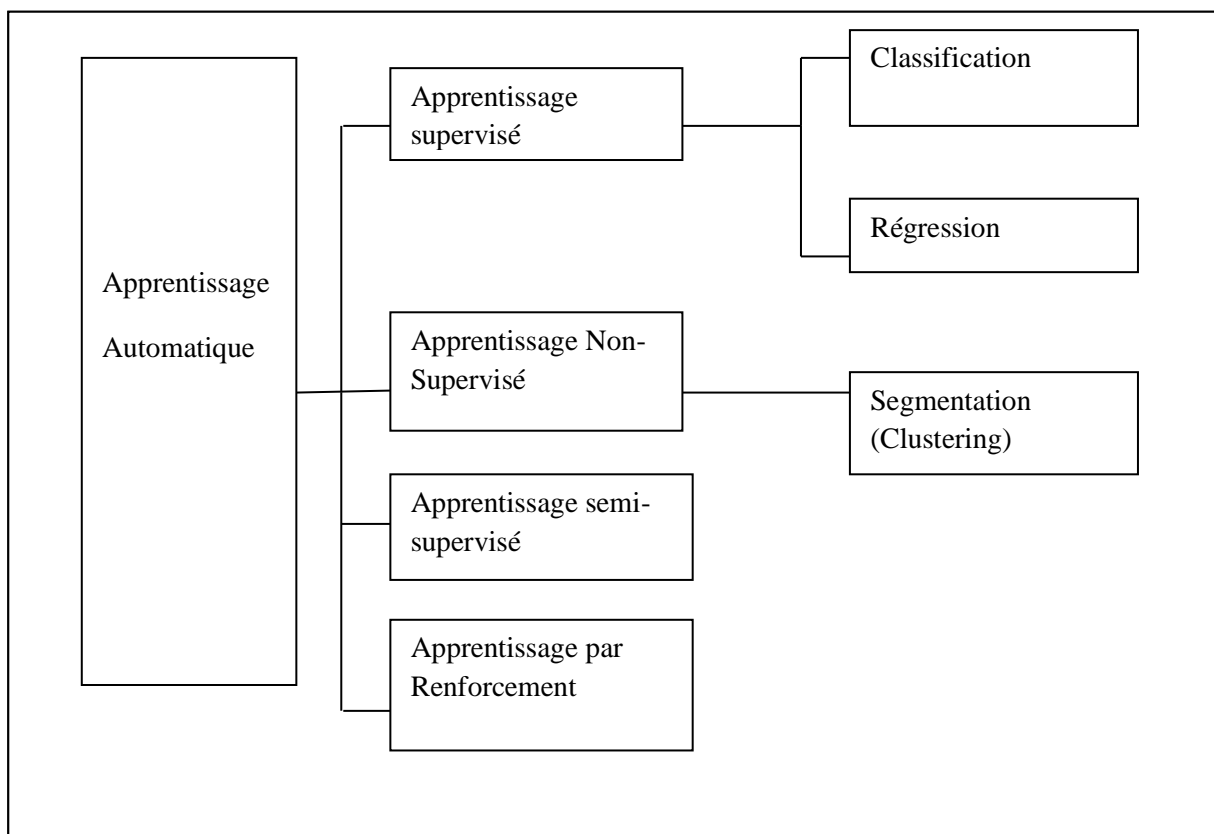


Figure II.2. Grandes classes ou types de l'apprentissage automatique.

II.4.5 Apprentissage profond (Deep Learning) :

Le Deep Learning, est une sous-branche du Machine Learning, qui se repose sur des réseaux neuronaux artificiels (Artificial Neural Networks ANN) comme structure, où

l'architecture de perceptrons multicouches (MLP MultiLayer Perceptron) est la plus basique (Goodfellow et al 2016).

Ce processus d'apprentissage est considéré comme profond car il est constitué d'une cascade de plusieurs couches réseaux de neurones : d'entrée, de sortie et masquées(Hinton et al 2006). Ces couches permettent de modéliser les liens linaires et non-linéaires entre les entrées et sortie de tels systèmes complexes.

Chaque couche renferme des éléments (neurones), qui convertissent les données d'entrée en informations qui pourrait être exploitées par la couche suivante pour une tâche prédictive particulière. Les calculs effectués par une première couche de "neurones" sont utilisés comme variables d'entrée de la deuxième couche et ainsi de suite.

Cette structure permet à une machine d'acquérir des connaissances grâce à son propre processus de traitement de données. Les algorithmes de l'apprentissage profond sont capables d'apprendre sans supervision humaine, possèdent la capacité d'extraire des caractéristiques significatives sans instruction explicite. Ils se distinguent également par leurs spécialité dans l'apprentissage de relations et liens précis, à partir d'ensembles de données volumineux et complexes, en utilisant des données à la fois non structurées et non étiquetées (LeCun et al 2015).

En Deep Learning, la profondeur du réseau est calculée en fonction du nombre de couches cachées dans le réseau neuronal et du nombre de couches en générale. La largeur du réseau est estimée en fonction du nombre de neurones pour chaque couche.

Parmi les domaines d'utilisation de l'apprentissage profond, on trouve dans la reconnaissance visuelle et tous problèmes de traitement d'images et vision par ordinateurs, ainsi que l'utilisation dans les systèmes composés de prédiction et modélisation d'événements aléatoires et complexes, utilisant des quantités énormes de données et de liens complexes et non linéaires. Les progrès de l'apprentissage profond ont été rendus possibles notamment grâce à l'augmentation de la puissance des ordinateurs et à la disponibilité de grandes bases de données. Parmi les différentes variantes de l'apprentissage profond (Deep Learning), nous présentons trois principales variantes :

II.4.5.1. Réseaux de neurones convolutifs

Les réseaux de neurones convolutifs (CNN Convolutionnal Neural Networks) sont un type d'algorithme de Deep Learning composé de couches de neurones et de filtres de convolution (LeCun, 1989). Ces modèles sont utilisés surtout dans le traitement d'images numériques. Il permet d'extraire des informations spatiales sur des images à l'aide de son architecture composée d'une succession de couches, à savoir (voir figure II.3) :

- Des couches de filtre de convolution : composées de filtres sous forme de fenêtre d'une taille inférieure à la taille de l'image traitée. Cette fenêtre est utilisée en translation sur l'image par des opérations de convolution mathématiques pour créer une carte d'attributs convolués (feature map).
- Des couches de Pooling (regroupement) : leurs rôle est de réduire la résolution des fenêtres convoluées, pour réduire la taille et quantité des données à traiter et faciliter l'extraction d'information précises et utiles. Cette partie fonctionne selon deux différents principes à savoir : le 'mean pooling' et le 'max pooling'.
- Les données traverseront ensuite la couche complètement connectée (fully-connected) qui est un Perceptron multi-couches (MLP) traditionnel. Une fonction d'activation est utilisée sur les données de sortie (output) afin d'ajouter la non-linéarité et permettre au réseau MLP de classifier les données.

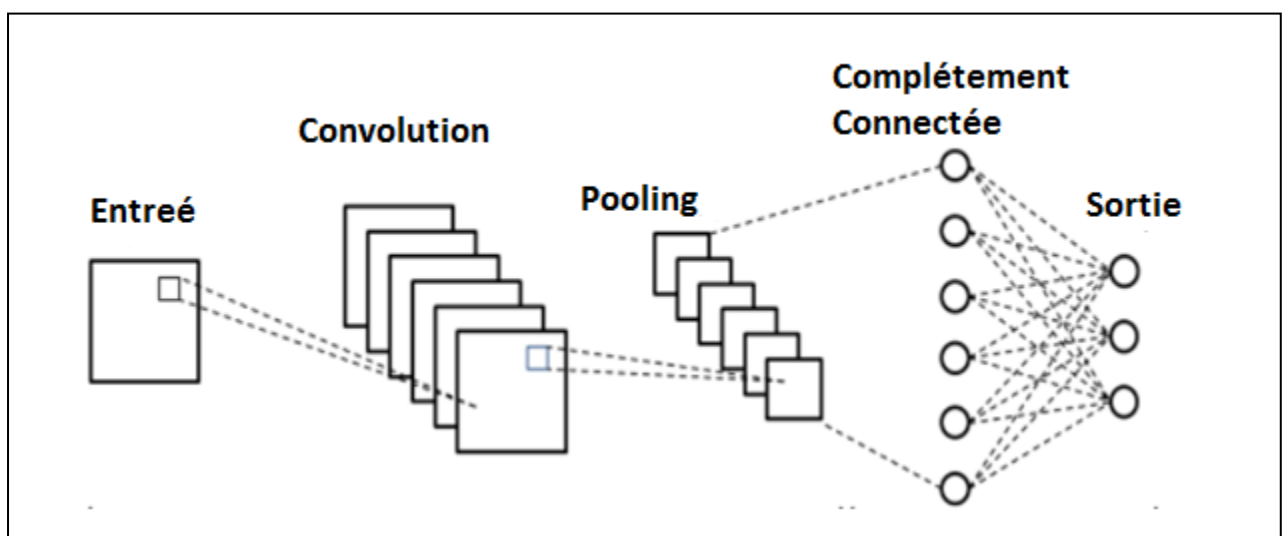


Figure II.3. Schéma de l'architecture de base d'un réseau CNN.

II.4.5.2. Réseaux de neurones récurrents (RNN) :

Les réseaux neuronaux récurrents connus sous le nom de RNN (Recurrent Neural Networks) (Pollack, 1990), sont des réseaux de neurones artificiels avec une ou plusieurs couches cachées récurrentes (Figure II.4). La différence entre une couche récurrente et une couche cachée régulière, est que les neurones d'une couche récurrente peuvent être connectés les uns aux autres ainsi qu'aux neurones de la couche suivante. En d'autres termes, les résultats des neurones de cette couche cachée de RNN sont transmis à la fois aux neurones de la couche suivante, pour donner les résultats de traitement d'une séquence et aux neurones de la même couche. Grâce à ce mécanisme, les RNN peuvent conserver la mémoire des informations du traitement précédent dans cette couche cachée à partir d'une séquence donnée, et de les transmettre au traitement de la séquence de données suivante. Ainsi la couche cachée peut utiliser le résultat de la séquence précédente dans le traitement de la séquence de données suivante. Cette procédure devient pratique lorsque les données à traiter par le réseau sont de nature séquentielle, comme des données de séries chronologiques.

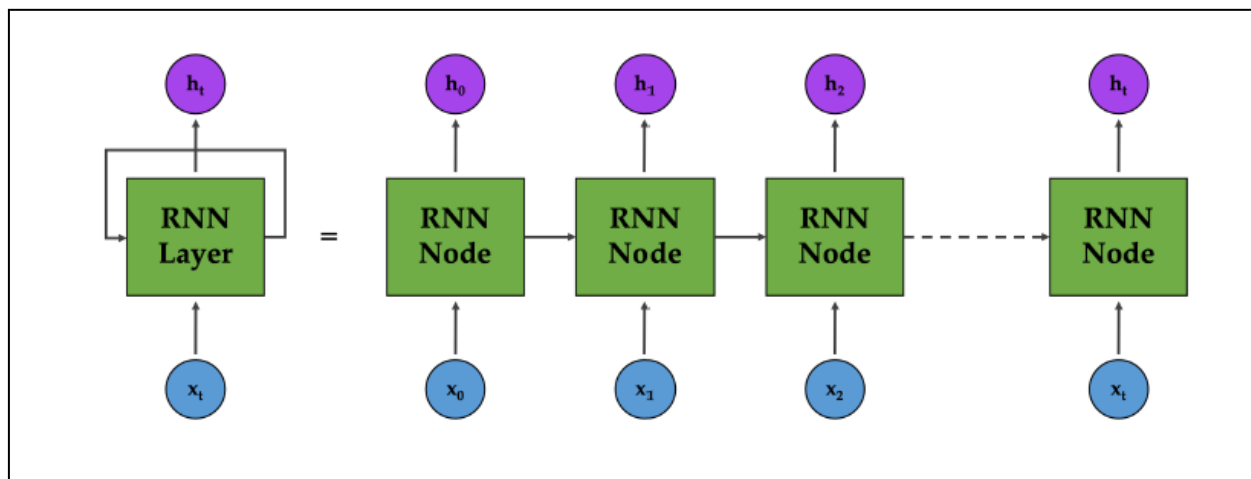


Figure II.4. Connexions récurrentes dans un réseau RNN.

II.4.5.3. Auto Encoders (AE) :

Les réseaux de neurones appelés auto-encodeurs (Rummelhart et al, 1985) sont utilisés pour diminuer la dimensionnalité des ensembles de données. Ils sont implémentés de manière non supervisée, pour générer uniquement une représentation de l'ensemble de données au sein de leurs neurones de couche cachée, également appelés vecteur latent.

Un AE apprend à réduire un ensemble de données dans un état de représentation, en utilisant le même ensemble de valeurs pour l'entrée et la sortie du réseau. Il apprend également à reconstruire l'échantillon de données dans sa forme originale à l'aide de la représentation apprise. La figure II.5 représente un schéma de l'architecture de l'Auto-Encodeur.

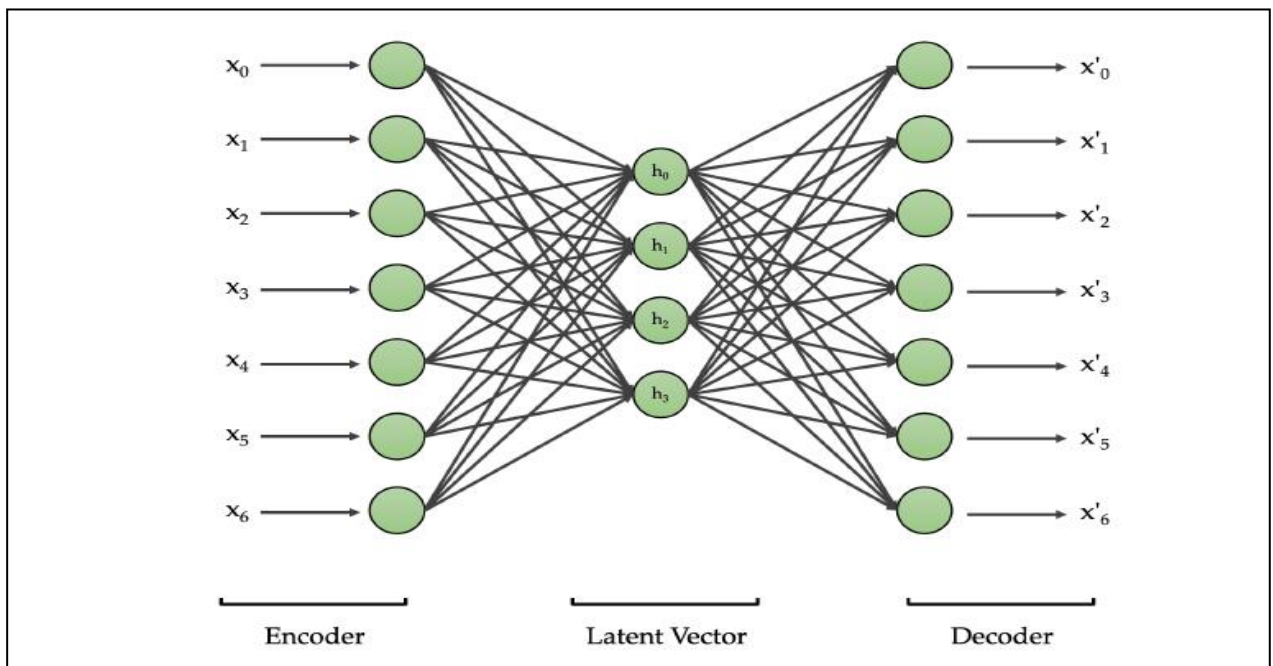


Figure II.5. Schéma de l'architecture de l'Auto-Encodeur.

II.5. Choix du model apprentissage automatique et modélisation du problème

La sélection de l'algorithme d'apprentissage automatique adéquat, pour résoudre un problème déterminé, est une étape primordiale avant de procéder à la modélisation et implémentation du modèle.

II.5.1. Choix du model d'apprentissage automatique

Etant donné qu'il existe plusieurs types d'algorithme d'apprentissage automatique, il est donc d'une importance capitale de savoir lequel utiliser. Ainsi, on choisira le type d'apprentissage à appliquer, en fonction de ce qui est envisageable pour la tâche souhaitée, et du type de données disponible ainsi que le traitement adéquat pour chaque type. Effectivement, le choix du type d'apprentissage et du modèle est basé sur les hypothèses formulées pour un problème spécifique.

L'utilisation et le choix d'un type d'algorithme d'apprentissage automatique est fortement dépendante de la nature du problème à résoudre (classification, régression, estimation de valeurs, etc.).

Il est donc essentiel d'avoir une certaine expertise des données disponibles et des contraintes liées au système (mémoire, puissance de calcul...), afin de sélectionner un modèle approprié pour la tâche en question. Il est crucial d'examiner à nouveau les hypothèses pour chaque problème envisagé et d'en déduire quel algorithme sera le plus efficace dans le cas considéré.

II.5.2. Modélisation d'un problème

Selon Goodfellow et al. (2016), un algorithme d'apprentissage automatique est un modèle statistique apprenant de ses expériences afin d'optimiser ses résultats. La nature de ces expériences et résultats diffère selon que le type du modèle choisi, soit l'apprentissage supervisé ou non supervisé, et son adaptabilité aux données. Formuler et modéliser un problème d'apprentissage automatique, revient à la description de la manière avec laquelle le modèle choisi doit traiter les données. L'exemple ou bien l'échantillon à traiter est représenté souvent sous forme d'un vecteur. On traite typiquement un vecteur $X \in R^n$ représenté par 'n' attributs ou propriétés différentes.

Le type de problème traité détermine le résultat obtenu en sortie du modèle d'apprentissage automatique. Cela peut être un résultat d'une classification binaire, ou multiclassées ou d'une régression pour le cas d'un problème d'apprentissage supervisé, et ça peut être d'autres catégories dans le cas d'autres types d'apprentissage automatique.

II.5.3. Cas d'apprentissage automatique supervisé

Nous nous intéresserons dans cette partie à la modélisation et formulation des problèmes d'apprentissage supervisé, vu que c'est le type d'apprentissage automatique le plus fréquemment employé.

Il s'agit de formuler les algorithmes et les méthodes, qui soient capables d'apprendre des modèles prédictifs à partir d'exemples étiquetés, et ce en observant dans leurs phase d'apprentissage plusieurs vecteurs aléatoires x et leurs valeur associé y afin d'apprendre à prédire y en fonction de x en modélisant leurs lien, qui est la fonction pour la quelle l'apprentissage est modélisé (Mehta et al 2019).

Le but de cette partie est de montrer et définir les concepts généraux, qui nous permettent de formuler ce type de problèmes, ainsi que les symbolisations et notions utilisées.

On peut formuler un problème d'apprentissage supervisé de la manière suivante: étant donné 'n' observations $\{x_1, x_2, \dots, x_n\}$, où chaque observation x_i est un élément de l'ensemble des observations X , et leurs étiquettes $\{y_1, y_2, \dots, y_n\}$ où chaque étiquette y_i appartient à l'ensemble des étiquettes Y , le but de l'apprentissage supervisé est de trouver une fonction $f : X \rightarrow Y$, qui modélise le lien entre les observations et les étiquettes, telle que : $f(x) \approx y$ pour toutes les paires $(x, y) \in X \times Y$. L'ensemble de $D = \{(x_i, y_i)\}_{i=1, \dots, n}$ forme le jeu d'apprentissage.

En apprentissage supervisé trois cas particuliers selon les valeurs de Y sont considérés:

- $Y = \mathbb{R}$: on parle d'un problème de régression ;
- $Y = \{0, 1\}$: il s'agit d'un problème de classification binaire, dans certains cas, il sera mathématiquement plus simple d'utiliser $Y = \{-1, 1\}$;
- $Y = \{1, 2, \dots, C\}$, $C > 2$: c'est un problème de classification multiclassés.

Dans de nombreuses situations du monde pratique où les données (observations) sont multidimensionnelles, on se ramènera au cas où $X = \mathbb{R}^p$. On dira alors que les observations (échantillons) x sont représentées par 'p' variables d'attributs.

Dans ce cas, les observations x sont représentées par une forme matricielle, avec la matrice des données $[X] \in \mathbb{R}^{n \times p}$, telle que l'élément $[X]_{ij} = x_j^i$ soit la j -ème attribut (caractéristique) de la i -ème observation (échantillon), est une matrice représentant les vecteurs de données (observations), avec 'n' qui est le nombre d'observations et 'p' le nombre d'attributs d'une observation x .

II.5.4. Décision dans le cas de classification

Dans le cas d'un problème de classification binaire, le modèle prédictif peut prendre la forme d'une fonction f à valeurs dans $\{0, 1\}$, qui nous donne le résultat de la classification, il y a certaines méthodes qui utilisent une fonction intermédiaire 'g' à valeurs réelles, qui attribue un score d'autant plus élevé à une observation « x » qu'elle est susceptible d'appartenir à la classe positive. On obtient alors f en seuillant g , tel que g est appelée fonction de décision qui donne le résultat de la classification de l'observation x .

Ainsi, on appelle fonction de décision, une fonction $g : X \rightarrow \mathbb{R}$ telle que :

$$f(x) = \begin{cases} 0 & \text{si et seulement si } g(x) \leq 0 \\ 1 & \text{si et seulement si } g(x) > 0 \end{cases} \quad (2.1)$$

Dans le cas de la classification multi-classe, la définition se généralise, on a alors 'C' fonctions de décision $g_c : X \rightarrow \mathbb{R}$, telles que

$$f(x) = \text{Argmax}_{c=1, \dots, C} g_c(x). \quad (2.2)$$

Donc le classifieur attribue à l'observation 'x' la classe qui présente la fonction de décision $g_c(x)$ avec les plus grand score.

Parmi les méthodes d'apprentissage supervisé utilisées dans le domaine de la classification, certaines permettent de résoudre directement des problèmes de classification multi-classe. Cependant, parfois on est confronté à des cas où on cherche à utiliser des algorithmes de classification initialement bi-classes et les élargir à une classification multiclassées. Pour cela il existe des méthodes (stratégies) qui permettent à tout algorithme de classification binaire d'être utilisé pour résoudre un problème de classification multiclassées à

C classes (Lorena et al 2008). C'est le cas des approches une-contre-toutes ou l'approche une-contre-une.

II.5.4.1. Classification Une-contre-toutes (ONE vs ALL) :

Étant donné un problème de classification multi-classe à C classes, on appelle une-contre-toutes ou 'one-versus-all', l'approche qui consiste à entraîner 'C' classifieurs binaires. Tel que tous les exemples de la classe 'C' sont utilisés par le C-ème de ces classifieurs comme étant des exemples positifs, tandis que toutes les autres observations sont considérées comme des exemples négatifs. Il se base sur la distinction des observations d'une 'classe c' par rapport à toutes les autres classes, pour apprendre une fonction de décision g_c qui est la probabilité que l'observation soit de la classe C.

Ainsi, chaque classifieur apprend à distinguer une classe de toutes les autres pour une observation donnée 'x'.

L'étiquette ou la classe d'un vecteur 'x' qu'est $f(x)$ est donnée par la fonction de décision $g_c(x)$ qui retourne le score le plus élevé :

$$f(x) = \arg \max g_c(x) \quad , \quad c = 1, \dots, C \quad (2.3)$$

II.5.4.2. Classification Une-contre-Une (ONE vs ONE) :

Dans le cas de problème de classification multi-classe à 'C classes', on appelle 'une-contre-une' ou 'one-versus-one', la variante qui consiste à créer 'C (C - 1) / 2' classifieurs binaires, séparant chacun d'eux entre deux classes spécifiques en ignorant toutes les autres classes (Lorena et al 2008).

Soit g_{ck} la fonction de décision du classifieur binaire, qui sépare la classe 'c' de la classe 'k'. L'étiquette de x est déterminée selon :

$$f(x) = \arg \max_{c=1, \dots, C} (\sum_{k \neq c} g_{ck}(x)) \quad (2.4)$$

Où x sera affecté à la classe 'c' qui affiche le score maximum.

Il est compliqué de déterminer laquelle de ces deux variantes est la plus efficace. Dans la réalité, le choix sera généralement influencé par des facteurs tels que la complexité algorithmique, la nature et l'équilibre des distributions des données des différentes classes, ainsi que les stratégies d'agrégation des sorties (Galar et al 2011). Est-ce qu'il est plus efficace

d'entraîner C modèles sur ' n ' observations, ou ' $C(C-1)/2$ ' modèles sur $2 \cdot \frac{n}{c}$ observations, en supposant les classes équilibrées, autrement dit que D contient sensiblement autant d'exemples de chaque classe. De plus, il est également important de considérer le nombre d'exemples disponibles pour chaque classe (la taille des classes) : si ce nombre est trop faible, il sera difficile d'obtenir un modèle de bonne qualité, et qui généralise bien sur les nouvelles données.

II.6. Préparation de données et fonctionnement en cas de la classification

Dans cette partie, nous décrivons en premier lieu, les étapes de préparations des données qui sont utilisées dans un algorithme d'apprentissage automatique. En deuxième lieu, nous montrons les étapes qui constituent le déroulement d'un algorithme d'apprentissage automatique dans le cas d'une classification.

II.6.1. Etapes de préparations des données :

Les étapes utilisées pour préparer les données sont résumées ainsi :

II.6.1.1. Acquisition des données

Cette étape fait référence aux processus de mesures de phénomènes physiques, et à l'enregistrement de ces mesures afin de les analyser et traiter pour en tirer des informations. Les signaux acquis par des capteurs tels que (radiomètres abord de satellites de télédétection, capteurs de température, capteurs de son...etc) sont convertis du domaine analogique au domaine numérique, puis enregistrés sur des supports numériques pour être disponible aux algorithmes d'apprentissage automatique, après leurs préparation.

II.6.1.2. Prétraitement

Les performances d'un classifieur d'apprentissage automatique dépendent aussi de la phase de prétraitement et préparation des données. Cette phase se compose d'une succession de traitements sur les données brutes acquises, afin d'en extraire de l'information. En général, les données ciblées sont celles utiles à la classification répondant aux objectifs du modèle d'apprentissage automatique construit. Cette phase de prétraitements sur les données brutes vise de nombreux objectifs comme :

- La réduction de la taille des données qui vont être introduites au classifieur, ce qui se traduit par un gain en temps de l'algorithme d'apprentissage, et rend la gestion et traitement des données plus facile et efficace.
- Éliminer l'information non pertinente, et les données bruitées et redondantes qui peuvent être une source de confusion pour l'algorithme d'apprentissage automatique ou classificateur, ce qui permettra une meilleure généralisation de l'algorithme.

Cette étape de prétraitement nécessite l'utilisation de certaines méthodes que nous citons ci-dessous :

i. Extractions des caractéristiques

L'application des techniques d'extractions des caractéristiques, permettent d'éliminer les redondances, et de ne conserver que l'information pertinente et significative pour le modèle. Cette phase nécessite l'intervention des experts du domaine étudié, pour leurs connaissances de données et pour une meilleures distinction entre données utiles ou non utiles.

ii. Sélections des caractéristiques

On a souvent le nombre de caractéristiques qui reste trop élevé, pour servir d'information d'entrée utiles d'un classificateur, même après l'application d'une technique de compression de données. L'objectif de cette phase de sélection de caractéristiques, est de trouver les caractéristiques pertinentes, qui vont contribuer à une bonne distinction entre deux éléments appartenant à deux classes hétérogènes, et un bon rapprochement entre les éléments d'un même ensemble, ce qui va enlever les données inutiles et réduire le temps de déroulement de l'apprentissage.

II.6.1.3. Conversion

Il s'agit de convertir les données dans le format utilisable par le classifieur. La représentation vectorielle est assez commune et plus souvent utilisée, notamment parce qu'elle permet de normaliser la modélisation des données. Tel que, les données et objets de classification sont représentés sous forme de vecteurs, dont chaque composante correspond à une caractéristique ou attribut de l'objet à classifier.

II.6.1.4. Post-traitement

Dans certains cas, on doit normaliser les données dans le format d'entrée, selon les caractéristiques du modèle d'apprentissage automatique utilisé et du phénomène à modéliser.

Après l'étape de traitement et préparation des données, pour les introduire à l'utilisation du modèle d'apprentissage automatique choisit, l'ensemble des données disponibles est divisé en deux ou trois parties appelées jeu de données (Datasets). Chaque partie est destinée à être utilisée dans une des étapes de déroulement et conception du model d'apprentissage automatique.

II.6.2. Etapes de fonctionnement de l'apprentissage automatique :

Pour la mise au point d'un algorithme basé sur l'apprentissage automatique, trois étapes sont réalisées.

II.6.2.1. Apprentissage

Le modèle d'apprentissage automatique est appliqué sur un ensemble de données d'apprentissage, qu'est étiquetées dans le cas d'apprentissage supervisé. Il est utilisé pour ajuster et optimiser les paramètres du modèle tels que : les poids et coefficients d'un classificateur, et pour permettre au modèle d'apprendre les liens existant entre les données d'entrée et les données de sortie qui sont les cibles (classes).

Dans la pratique, Le jeu de données d'apprentissage se compose souvent de paires d'un vecteur d'entrée représentant l'objet (échantillon) à classifier et d'un vecteur de sortie correspondant, qui est la variable de réponse appelée cible (ou étiquette). Le modèle est exécuté avec un jeu de données d'apprentissage et produit un résultat, qui est ensuite comparé à la cible. Les paramètres du modèle sont ajustés en fonction des résultats de la comparaison et de l'algorithme d'apprentissage spécifique utilisé.

L'objectif est de produire un modèle entraîné (ajusté), qui se généralise bien aux nouvelles données inconnues. Les nouveaux exemples des jeux de données de validation et de test sont utilisés pour évaluer le modèle ajusté.

II.6.2.2. Etape de validation

Le modèle est utilisé dans cette étape sur des données de validation, qui sont un ensemble de données utilisés pour régler les hyper paramètres c'est-à-dire l'architecture du model, et les paramètres caractéristiques d'un classifieur. Un exemple d'hyper paramètres à

réglé dans le cas des réseaux de neurones artificiels est le nombre de couches intérieures (cachées).

Il est important de disposer d'un jeu de données de validation, en plus des jeux de données d'apprentissage et de test, afin d'éviter le sur-apprentissage, lorsqu'un paramètre du classifieur doit être ajusté ou optimisé. Le jeu de données de validation fonctionne comme un hybride : ce sont des données d'entraînement utilisées pour les tests, dans le cas de l'optimisation de paramètres.

II.6.2.3. Etape de test

Une fois le modèle appris, et ses paramètres ajustés dans les étapes précédentes, on réalise des tests pour évaluer les performances sur un ensemble indépendant de données, appelé données de validation. Cette phase permet de vérifier le pouvoir de généralisation du classifieur, c'est à dire sa capacité à obtenir de bons résultats en l'appliquant à des données sur les quels il n'a pas été entraîné, et sur n'importe quel ensemble de données provenant de la même distribution.

Le terme « jeu de validation » est parfois utilisé au lieu de « jeu de test » dans certains cas, où les données d'origine ont été divisées en deux sous-ensembles seulement, ainsi l'ensemble de données de test peut être appelé ensemble de validation. Le partage de ces ensembles dépend beaucoup du type de problème et des données disponibles.

Dans les cas où l'on ne dispose pas de suffisamment de données, pour les diviser en différents ensembles (apprentissage, validation et test), une technique appelée 'validation croisée' est utilisée, pour générer plusieurs paires de sous-ensembles (entraînement / test). Cette technique est coûteuse en temps de calcul, mais permet d'obtenir une bonne estimation de l'erreur de test, tout en conservant suffisamment d'exemples pour l'entraînement du modèle.

II.6.3. Validation croisée

La validation croisée est une technique qui permet de tester un modèle d'apprentissage, et d'évaluer ses performances. Ceci est dans le cas où les tailles de données à disposition ne sont pas suffisantes pour effectuer le découpage en différents ensembles de

données nécessaires d'apprentissage et de test, toute en ayant une taille suffisante de données pour obtenir un modèle appris qui généralise bien.

Cette méthode a plusieurs variantes, la plus connue est la méthode « k-Fold » avec typiquement $k \in [4, 10]$ qui constitue les valeurs les plus souvent utilisées de k . Si l'on a une base d'apprentissage A_p contenant 'p' éléments ou observations (Refaeilzadeh et al 2009) :

$$A_p = x_1, \dots, x_p .$$

La validation croisée s'effectue selon les cinq étapes suivantes:

- Découper l'ensemble des exemples de données en k sous-ensembles disjoints de taille p/k .
- Entraîner le modèle sur les $k-1$ sous-ensembles en les considérant comme données d'apprentissage.
- Calculer l'erreur sur la *kième* partie en l'utilisant comme ensemble de validation.
- Répéter le processus K fois en changeant à chaque fois l'ensemble de validation.
- Obtenir l'erreur finale en calculant la moyenne des K erreurs précédentes.

La mise en œuvre de la validation croisée est simple et elle exploite toutes les données disponibles. Elle permet d'évaluer l'erreur de généralisation. Ce qui permet d'éviter ou de réduire le risque de sur-apprentissage, qui est une notion que nous définissons dans la partie suivante.

II.7 .Notions de généralisation et sur-apprentissage

Nous définissons dans cette partie quelques notions de base utilisées pour décrire les modèles d'apprentissage automatique :

II.7.1. Généralisation

Le terme « généralisation » fait référence à la capacité d'un modèle d'apprentissage automatique à faire des prédictions précises concernant de nouvelles données qui n'ont pas été utilisées pendant sa phase d'apprentissage et de construction. L'évaluation d'un algorithme d'apprentissage automatique sur les données avec lesquelles il est entraîné ne donne pas suffisamment d'informations sur son éventuel comportement envers de nouvelles données non

rencontrées pendant la phase d'apprentissage, c'est-à-dire sa capacité de généralisation. La généralisation est cruciale car les données que nous utilisons dans l'apprentissage ne sont qu'un échantillon des données sur le phénomène étudié et peuvent être incomplètes et contenir du bruit, ce qui les empêche de représenter entièrement le phénomène.

En apprentissage automatique, lorsque nous parlons de la capacité d'un modèle à généraliser sur de nouvelles données, les termes et notions de sur-apprentissage et le sous-apprentissage sont utilisés.

II.7.2. Sur-apprentissage :

Un modèle en sur-apprentissage est généralement un modèle trop complexe, qui s'adapte trop aux données d'apprentissage, et capture donc tous les aspects caractérisant ces données y compris leurs fluctuations aléatoires et bruit en les apprenants en tant que concept. Ce qui nous donne de très grandes erreurs lorsque les données qui n'appartiennent pas aux données d'apprentissage doivent être classées. Le sur-apprentissage (over-fitting en Anglais) apparaît lorsque l'erreur du modèle sur les données de test est significativement plus élevée que sur les données d'apprentissage.

II.7.3. Sous-apprentissage

Lorsque le modèle prédictif basé sur l'apprentissage automatique n'arrive même pas à capturer ou modéliser les corrélations et liens entre entrées et sorties des données d'apprentissage. Il ne pourra pas bien se généraliser non plus sur les nouvelles données qu'il n'a pas encore rencontrées. Le modèle ne sera fiable car les erreurs de prédictions seront grandes. Dans ce cas de figure, on dit que le modèle souffre de sous-apprentissage (Underfitting). On dit également qu'il souffre d'un grand Bias (biais). Il est crucial de s'assurer que les modèles de prédiction ou de classification que nous souhaitons mettre en œuvre ne présentent ni sur-apprentissage ni sous-apprentissage.

La figure II.6 illustre la différence entre un bon modèle d'apprentissage automatique et des modèles en sur-apprentissage et sous-apprentissage, respectivement.

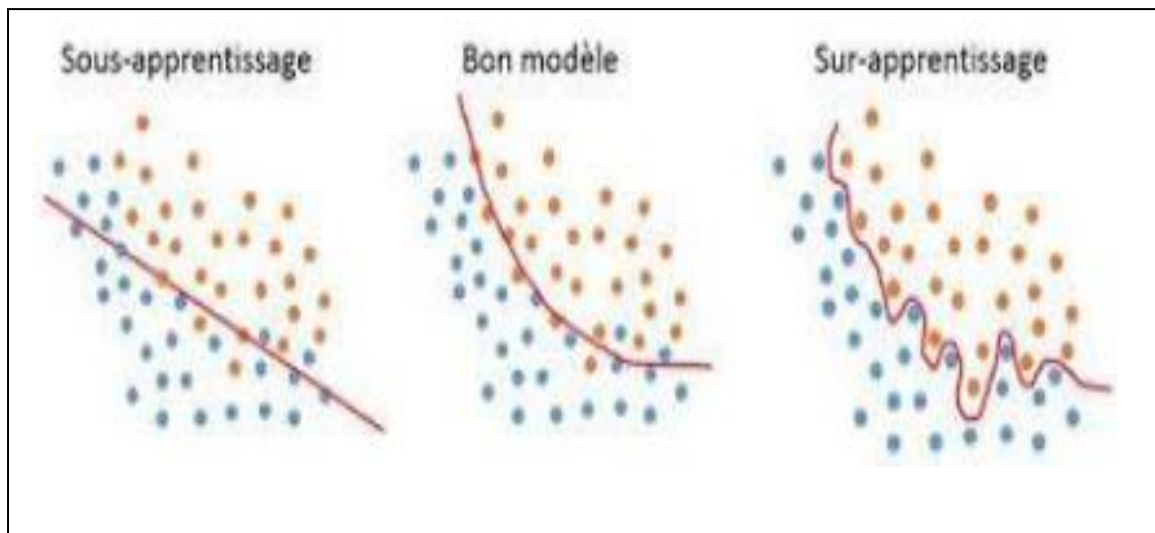


Figure II.6 : Illustration de modèle en sous-apprentissage et sur-apprentissage.

II.8. Quelques modèles d'apprentissage automatique

Nous nous focalisons dans cette partie sur la définition de quelques modèles d'apprentissage automatique supervisé, les plus pertinents et les plus utilisés en classification.

II.8.1. Réseaux de neurones artificiels

Les réseaux de neurones sont issus de l'imitation de la structure neurophysiologique du cerveau, et leur application a émergée, depuis l'introduction du model du perceptron par (Rosenblatt,1958). Le réseau de neurones artificiels contient un ensemble de neurones artificiels fortement interconnectés, dans le but d'imiter certaines fonctions du cerveau humain, tel que la mémorisation par association, l'apprentissage par exemple et la modélisation des liens.

Ces réseaux de neurones sont des systèmes capables de modifier leur structure et paramètres internes en fonction des données qu'ils traitent. Ils sont particulièrement adaptés à la résolution de problèmes de type non linéaire, souvent rencontrés dans la réalité.

Un neurone formel est l'unité élémentaire d'un système modélisé par un réseau de neurones artificiels (figure II.7). Il est à la réception de signaux provenant d'autres neurones de la couche précédente du réseau, ayant des valeurs numériques x_1, x_2, \dots, x_n , auxquels il calcule une sortie y . Chaque entrée x_i a un poids associé w_i , qui est attribué en fonction de son importance relative par rapport aux autres entrées. La valeur d'entrée x du neurone correspond

à la somme pondérée de ses entrées, qui sont les sorties des neurones de la couche précédente en y ajoutant une autre entrée ayant un poids b appelé biais (Figure II.7). Un neurone formel produit un signal de sortie y en appliquant une fonction f appelée fonction d'activation du neurone, cette sortie ' y ' qui sera transmise à d'autres neurones de la couche suivante du réseau.

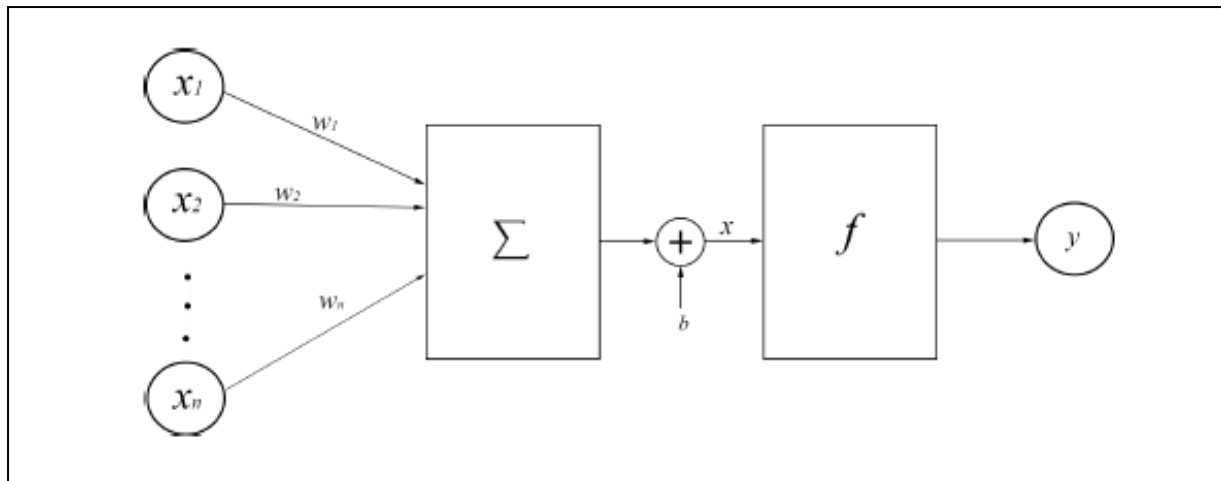


Figure II.7 Schéma d'un neurone formel avec les signaux d'entrée et sortie.

La sortie ' y ' du neurone est calculée selon la formule suivante :

$$y = f(\sum_{i=1}^n w_i x_i + b) \quad (2.5)$$

En classification, les réseaux de neurones ont permis d'introduire la non-linéarité dans la séparation entre les classes grâce au choix de la fonction d'activation ' f ', qui introduit la non-linéarité dans le signal passant de son entrée à sa sortie, permettant ainsi la modélisation des liens non-linéaires entre les données d'entrée et sortie. Le réseau de neurones à propagation avant ou (feed-forward neural network), était le premier et le plus simple réseau de neurones artificiel mis au point. Il contient plusieurs neurones disposés en couches, tel que les neurones des couches adjacentes ont des connexions entre eux. Toutes ces connexions ont des poids qui leur sont associés.

Un réseau de neurones Feedforward peut être constitué de trois types de neurones :

- **Neurones d'entrée** : les neurones d'entrée fournissent des informations de l'environnement extérieur au réseau, ou bien permettent de faire rentrer des données des systèmes à modéliser vers le réseau, et construisent ensemble une « couche

d'entrée ». Aucun calcul n'est effectué dans les nœuds d'entrée. Ils transmettent simplement les informations aux nœuds cachés.

- **Neurones cachés** : les neurones cachés n'ont pas de connexion directe avec le monde extérieur. Ils effectuent des calculs et transfèrent des informations des neurones d'entrée aux neurones de sortie. Un ensemble de neurones cachés forme une « couche cachée ».
- **Neurones de sortie** : les neurones de sortie construisent ensemble une « couche de sortie ». Ils sont responsables des calculs et du transfert des informations du réseau vers le monde extérieur.

Dans un réseau feedforward neural networks , les informations ne se déplacent que dans un seul sens, d'où son nom, en partant des neurones d'entrée, passant par les neurones cachés et en arrivant aux neurones de sortie(Goodfellow et al 2016). Ces réseaux peuvent être 'mono-couche' ou 'multi-couches', selon leurs configurations et le nombre de couches cachées de neurones dont ils disposent.

Dans la pluparts des applications actuelles, les perceptrons multi-couches MLP (Multilayer Perceptron) sont plus utiles que les perceptrons mono-couches. Le schéma de la figure II.8 illustre les deux types de réseaux de neurones à propagation avant, à savoir les 'mono-couche' et 'multi-couches'.

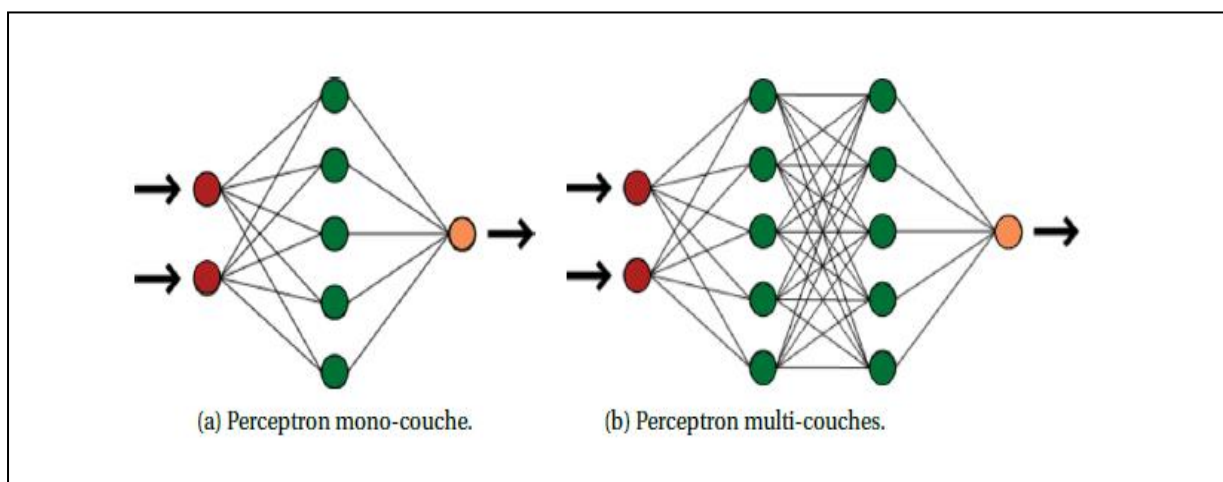


Figure II.8 : Représentation des réseaux de neurones mono-couche et multicouches (exemple d'un réseau MLP).

En se basant sur l'équation (2.5) de fonctionnement du perceptron formel précédente, nous pouvons décrire la sortie d'un neurone dans le réseau, tel que la sortie d'un neurone 'j' de la couche 'k' d'un réseau de neurones sera donnée comme suit :

$$y_j^k = f(\sum_{i=1}^n w_{ij}^{(k-1)} x_i + b) \quad (2.6)$$

Où :

y_j^k : La sortie du neurone 'j' de la couche 'k' .

f : La fonction d'activation de ce neurone 'j' de la couche 'k' .

$w_{ij}^{(k-1)}$: Les poids des liaisons entre les sorties des neurones 'i' de la couche 'k-1' et l'entrée du neurone 'j' de la couche 'k' .

x_i : Sortie du neurone 'i' de la couche 'k-1'.

b : biais du neurone 'j' .

n : nombre de neurone de la couche (k—1)

Chaque nœud dans le réseau de neurones artificiels, a une fonction d'activation f , par laquelle il transforme sa propre entrée globale en sortie, qu'il va transférer vers les entrées des neurones de la couche suivante du réseau de neurones.

Un réseau de neurones peut être défini et caractérisé par les paramètres suivants :

r : le nombre des neurones sur la couche d'entrée.

s : le nombre des neurones sur la couche de sortie.

l : le nombre total des couches du réseau de neurones.

n_k : le nombre des neurones sur la kème couche avec $0 \leq k \leq l$.

II.8.2. Algorithmes des K plus proches voisins (K-PPV)

L'algorithme des K plus proches voisins (Cover et Hart,1967) est un algorithme d'apprentissage supervisé. Il est utilisé donc avec l'ensemble de données $D = \{x_i, y_i\}_{i \in [1, N]}$ constituant de variables à classifier et leurs étiquettes.

Il s'agit d'un algorithme de classification, les sorties y_i correspondent alors à des classes possibles et non pas des valeurs continues.

Le principe de cet algorithme est que pour chaque nouvelle entrée vecteur 'x' à classifier, il mesure la distance entre 'x' et les entrées x_i déjà classifiées pour tout $i \in [1, N]$. Il sélectionne ensuite les K éléments les plus proches (distance plus faible) de l'entrée 'x'. La classe prédite par l'algorithme correspond alors à la classe la plus fréquente dans les K plus proches voisins sélectionnés du vecteur 'x'.

Différentes mesures de distance (euclidienne, Minkowski, Manhattan...) sont utilisées pour cet algorithme en fonction du problème étudié et ses caractéristiques.

Cet algorithme peut se montrer relativement robuste dans le cas où la taille des données 'apprentissage est suffisante. Le choix de la valeur du paramètre K dans l'algorithme des 'K plus proches voisins ' est aussi d'une influence capitale sur les résultats de classification de cet algorithme :

- K trop petit : des éléments aberrants pourront influencer plus facilement la prédiction. La généralisation du modèle ainsi développé sur de nouveaux éléments sera donc moins bonne, ce qui donne le problème de sur-apprentissage (overfitting).
- K trop grand : Le modèle prendra en compte beaucoup d'échantillons des données trop éloignées et la classe majoritaire sera prédite trop souvent. Ce qui donne ainsi le problème de sous apprentissage.

Pour le bon fonctionnement de cet algorithme, une taille suffisante de données d'entraînement est nécessaire afin qu'il y aura toujours des exemples à classe connues proches des exemples à classifier.

La figure II.9 illustre un exemple du processus de fonctionnement d'une classification utilisant l'algorithme des 'k voisins les plus proches KNN (K-Nearest Neighbours).

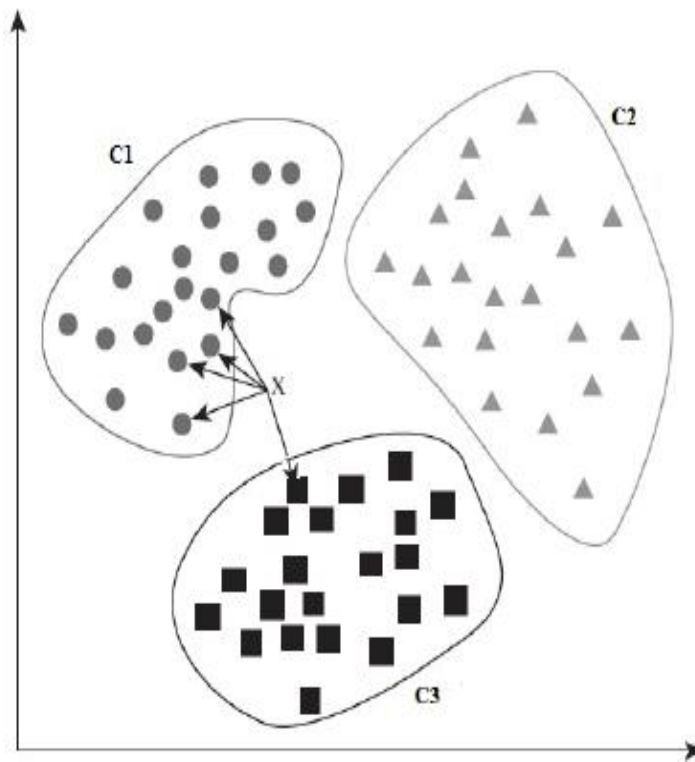


Figure II.9 : Exemple d'une classification par k-plus proches voisin(KNN).

Dans cet exemple, nous avons des échantillons qui ont été classifiés est affectés à trois classes (C1, C2, C3). L'objectif de l'algorithme de classification (KNN) ici est d'affecter le nouvel échantillon 'x' à une de ces trois classes. La distance Euclidienne est utilisée par l'algorithme ici pour calculer le voisinage du nouvel échantillon et une valeur de (k = 5 voisins) est choisit. Sur les cinq plus proches voisins de 'x' comme le montre le schéma (Figure II.9), quatre appartiennent à la classe C1 et un seul à la classe C3, l'échantillon 'x' est donc attribué à C1 en tant que classe à laquelle appartient la majorité de ses voisins.

II.8.3. Algorithme Machines à Vecteurs Support :

L'algorithme (SVM : Support Vector Machines) a pour but de trouver l'hyperplan séparateur optimal (OSH) entre deux classes en maximisant la marge (m) entre les deux échantillons les plus proches de ces classes se trouvant sur les frontières (Cortes et Vapnik, 1995). Ces points sont appelés vecteurs de supports (Voir Figure II.10).

En considérant un ensemble de données à classer comme $\{x_i, y_i\}, i = 1, \dots, N,$

$x_i \in R^d$, et $y_i \in \{1, -1\}$ l'étiquette de classe des échantillons x_i .

La frontière de décision pour une séparation linéaire des deux classes, illustrée par la (Figure II.10), est définie par :

$$W^T X + b = 0 \tag{2.7}$$

Où W est le vecteur normal à l'hyperplan de séparation entre les deux classes, $(|b|/\|W\|)$ est la distance perpendiculaire de l'hyperplan à l'origine, et $\|W\|$ est la norme euclidienne de W .

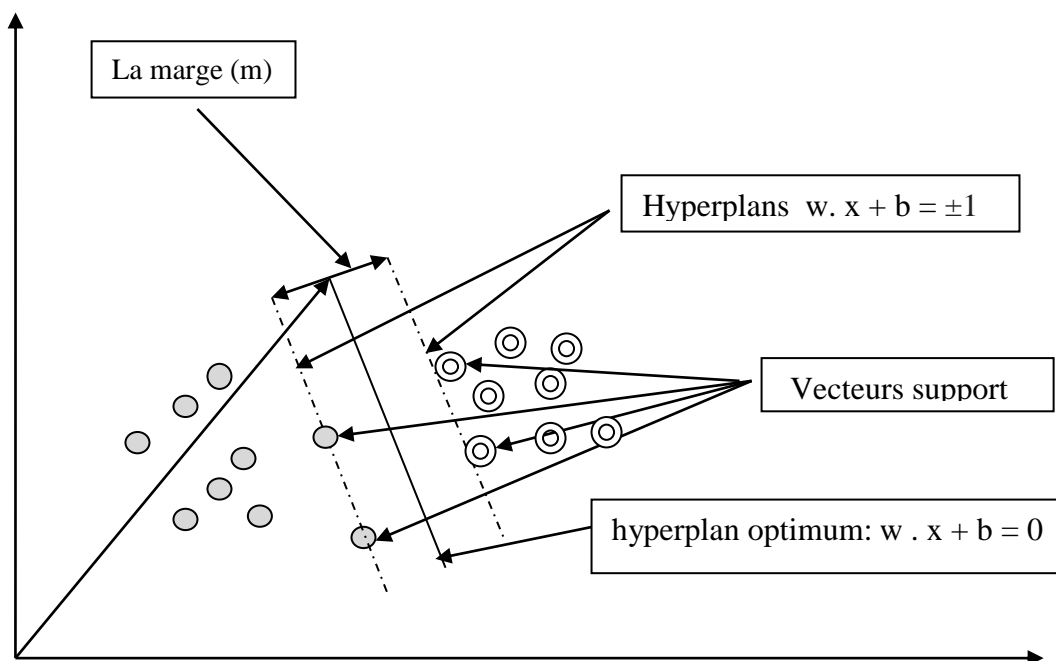


Figure II.10 : principes de marge et hyperplans séparateurs de SVM.

L'algorithme de la machine à vecteurs support recherche l'hyperplan séparateur optimal donnant la plus grande marge entre les deux classes.

Les conditions des deux classes délimitées par ce classificateur peuvent être combinées en un ensemble d'inégalités et ainsi, la frontière de décision est donnée par :

$$y_i(W^T x_i + b) \geq +1 \quad \text{pour } i=1, \dots, N \tag{2.8}$$

La distance entre les vecteurs séparés des deux classes constitue la marge (m) et est donnée par : $m = 2/\|W\|$. Par conséquent, SVM recherche la marge maximale en maximisant $2/\|W\|$ ce qui revient à minimiser $\|W\|^2/2$ sous contraintes données par l'expression (2.8).

Pour les problèmes de points non linéairement séparables, SVM utilise une fonction qui gère le compromis entre les erreurs de classification par rapport au taux de maximisation de la marge. Par conséquent, une variable de relâchement des contraintes ou "slack" $\xi_i \geq 0$ est introduite pour chaque exemple d'apprentissage, afin qu'elle fixe une marge d'erreur tolérée sur la séparation des frontières des deux classes, pour classer tous les échantillons de données. La résolution du problème d'optimisation qui résout cet élargissement de la marge entre les classes devient :

$$\text{Min} \left(\|W\|^2/2 + C \cdot \sum_i \xi_i \right) \quad (2.9)$$

Soumis à la contrainte :

$$y_i(W^T x_i + b) \geq 1 - \xi_i \text{ Avec } \xi_i \geq 0 \quad (2.10)$$

Où C est le paramètre qui contrôle le compromis entre l'erreur tolérée dans la classification et la séparation des échantillons et l'élargissement de la marge, et $(C \cdot \sum_i \xi_i)$ est un terme de pénalité.

Le principe de SVM repose sur une séparation linéaire des données dans un espace caractéristique de plus grande dimension, dans le cas où les vecteurs des deux classes ne sont pas linéairement séparables dans l'espace d'origine, et ce en représentant les données dans ce nouvel espace de dimension supérieure (éventuellement dimension infini) R^D à l'aide d'une fonction φ :

$$\varphi: R^d \longrightarrow R^D \text{ avec } : D \gg d.$$

D est appelé le nouvel espace de caractéristiques défini par la fonction φ . L'algorithme de marge maximale peut construire un hyperplan de séparation dans l'espace de caractéristiques

de dimension supérieure D, en évaluant les produits internes (scalaires) entre les vecteurs x et x_i dans le nouvel espace fonctionnalité de la forme : $\varphi(x_i)^T \cdot \varphi(x)$

Donc, la fonction K donne le produit scalaire de ces vecteurs, dans le nouvel espace de fonctionnalité de dimension supérieure, sans avoir à calculer les vecteurs de transfert $\varphi(x_i)$ vers ce nouvel espace de caractéristiques. Cette fonction noyau K (Kernel function) est définie par :

$$k(x_i, x_j) = K_{i,j} = \varphi(x_i)^T \cdot \varphi(x_j) \quad (2.11)$$

Avec : $W = \sum_i \alpha_i y_i \varphi(x_i)$, ainsi la fonction d'optimisation équivalente L (lagrangienne) à maximiser dans le nouvel espace, afin de maximiser la marge de classification (m) est :

$$L = -\frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \varphi(x_i) \varphi(x_j) + \sum_i \alpha_i = -\frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j K(x_i, x_j) + \sum_i \alpha_i \quad (2.12)$$

Ainsi minimiser la quantité de l'équation (2.9), revient à rechercher le maximum de L, équation (2.12) en fonction de α_i , sous contraintes :

$$C \geq \alpha_i \geq 0 \quad , \quad \sum \alpha_i y_i = 0 \quad \text{voir (Vapnik et Cortes ,1995).}$$

La fonction de décision qui donne les résultats de classification

$$F(x) = W^T \varphi(x) + b = \sum_i \alpha_i y_i \varphi(x_i) \varphi(x) + b = \sum_i \alpha_i y_i K(x_i, x) + b \quad (2.13)$$

Où :

x_i : est le vecteur support(support vector) , et x : le vecteur d'entrée à classifier , et (α_i , y_i) sont des multiplicateurs de Lagrange .

$W = \sum_i \alpha_i y_i \varphi(x_i)$ représente une combinaison de vecteurs supports, obtenue après vérification que la contrainte satisfait la condition de Karush Kuhn Tucker (KKT) (Vapnik et Cortes ,1995).

II.8.4. Arbres de décision

Un arbre de décision est un algorithme d'apprentissage supervisé utilisés à la fois pour les tâches de classification et de régression. Il construit une structure arborescente de type organigramme, il commence par des nœuds racine qui n'ont aucune branche rentrante, ils peuvent être considérés comme entrée de données. Ensuite chaque nœud interne désigne un test sur un attribut ou paramètre de données d'entrée, chaque branche représente un résultat du test et chaque nœud feuille (nœud terminal) contient une étiquette de classe, tel que les nœuds feuilles représentent à la fin les différents résultats ou classes possibles. Cet algorithme est construit en divisant les données d'entraînement en sous-ensembles, en fonction des valeurs des attributs et les seuils des tests aux nœuds, jusqu'à ce qu'un critère d'arrêt soit satisfait, tel que la profondeur maximale de l'arborescence. Parmi les variantes de cet algorithme nous citons l'algorithme CART de (Brieman et al 1984).

II.8.5. Méthodes d'ensemble (ensemblistes) :

Ces techniques sont des méthodes qui combinent plusieurs modèles ou algorithmes d'apprentissage automatique de base dans un même modèle prédictif, avec l'objectif d'améliorer la généralisation et la robustesse de ces modèles et de maximiser leurs performances. En effet, cette approche s'appuie sur l'idée que la moyenne des résultats d'un ensemble de classifieurs combinés, est statistiquement plus fiable que les résultats d'un seul classifieur.

Par conséquent, les méthodes d'ensemble permettent de créer une collection de prédicteurs et d'agrèger l'ensemble de leurs prédictions, bénéficiant ainsi des points forts de chacun des composants et essayant de limiter leurs points faibles.

Les méthodes combinées ou d'ensemble ont connus du succès dans une variété de problèmes. Cependant, il existe certains inconvénients avec ces techniques. Cela inclut le besoin d'espace mémoire supplémentaire et les coûts de calcul en combinant plusieurs classifieurs.

Ces méthodes d'ensemble utilisées dans l'apprentissage automatique, peuvent être divisées en deux catégories, selon la manière utilisée pour la combinaison des apprenants et la construction des modèles : les méthodes d'ensemble parallèles, comme Bagging (Brieman 1996) et les méthodes d'ensemble séquentielles comme l'algorithme 'Boosting' (Schapire 1990).

II.8.5.1. Méthodes d'ensemble parallèles

Dans le cas de ces méthodes, les modèles et classifieurs de base comme (les arbres de décision) sont générés de façon indépendante l'un de l'autre et en parallèle, comme l'exemple des forêts aléatoires (Random Forests RF) de (Breiman 2001).

La motivation et objectif derrière l'utilisation de ces méthodes est qu'elles peuvent réduire l'erreur de prédiction de manière remarquable en réduisant la variance, et ce grâce au fait que chaque classifieur est généré indépendamment sur l'ensemble de données.

L'algorithme de Bagging (Bootstrap aggregating) est une approche de construction d'algorithmes d'ensemble, qui génère les différents classifieurs composant la méthode, en utilisant différents sous-ensembles de données de l'ensemble d'apprentissage avec un même algorithme d'apprentissage, pour construire les différents prédicteurs composant l'ensemble.

L'algorithme tire des instances (échantillons) aléatoires 't' de l'ensemble de données avec remplacement (remise), ce qui fait que les instances tirées pour construire un prédicteur contiendront des redondances et des omissions par rapport à l'ensemble d'apprentissage initial, Bagging construit ainsi le sous-ensemble d'apprentissage appelé 'bootstrap' de chacun des apprenants composant la méthode d'ensemble, et chaque cycle de ce processus aboutit à un classifieur. Après la construction de plusieurs classifieurs, les sorties de tous ces classifieurs sont combinés pour produire la prédiction finale de la méthode.

II.8.5.2. Méthodes d'ensemble séquentielles

Pour ces méthodes comme le Boosting (Schapire, 1990), les classifieurs de base sont générés de manière séquentielle et dépendante, à l'inverse des méthodes parallèles. À chaque fois qu'un classifieur de base est entraîné, les exemples mal classifiés dans l'itération précédente sont pondérés avec un poids plus élevé pour que le classifieur de l'itération présente se focalise sur eux, dans le but que les erreurs des modèles précédents soient corrigés dans les itérations qui suivent. Ceci devra améliorer la performance globale.

Après plusieurs cycles, la prédiction finale est réalisée selon un vote pondéré des prédictions de chaque classifieur. Ainsi, le poids de décision qu'aura chaque classifieur dans le résultat final est proportionnel à la précision de chaque classifieur sur son ensemble d'apprentissage. Parmi ces méthodes on trouve l'algorithme AdaBoost.

➤ **AdaBoost :**

L'algorithme AdaBoost ou (Adaptive Boosting) est un algorithme qui utilise des classifieurs de base qui devront être relativement faibles pour construire un ensemble puissant et plus précis (Freund et Schapire 1997).

À la fin, le modèle génère une classification combinée, qui est la moyenne pondérée des résultats de classification des différents classifieurs de base. En effet, pour chaque apprenant de base, AdaBoost lui attribue un poids qui va compter dans le résultat final de la combinaison des apprenants, en fonction de la performance et précision individuelle de cet apprenant durant son itération. L'idée est que les classifieurs devront se focaliser sur les instances qui ont été difficiles à classer correctement par ceux des itérations qui les ont précédés. Tout algorithme d'apprentissage automatique peut être utilisé comme classifieur de base ou (apprenant faible) de AdaBoost, s'il peut accepter l'attribution des poids et coefficients sur l'ensemble d'entraînement et de satisfaire la condition sur sa précision.

La description de l'algorithme de AdaBoost, et le détail sur les étapes de fonctionnement ainsi que la théorie de cet algorithme sont décrits dans le (chapitre 4).

La figure II.11 illustre les schémas de base des principales méthodes d'ensemble.

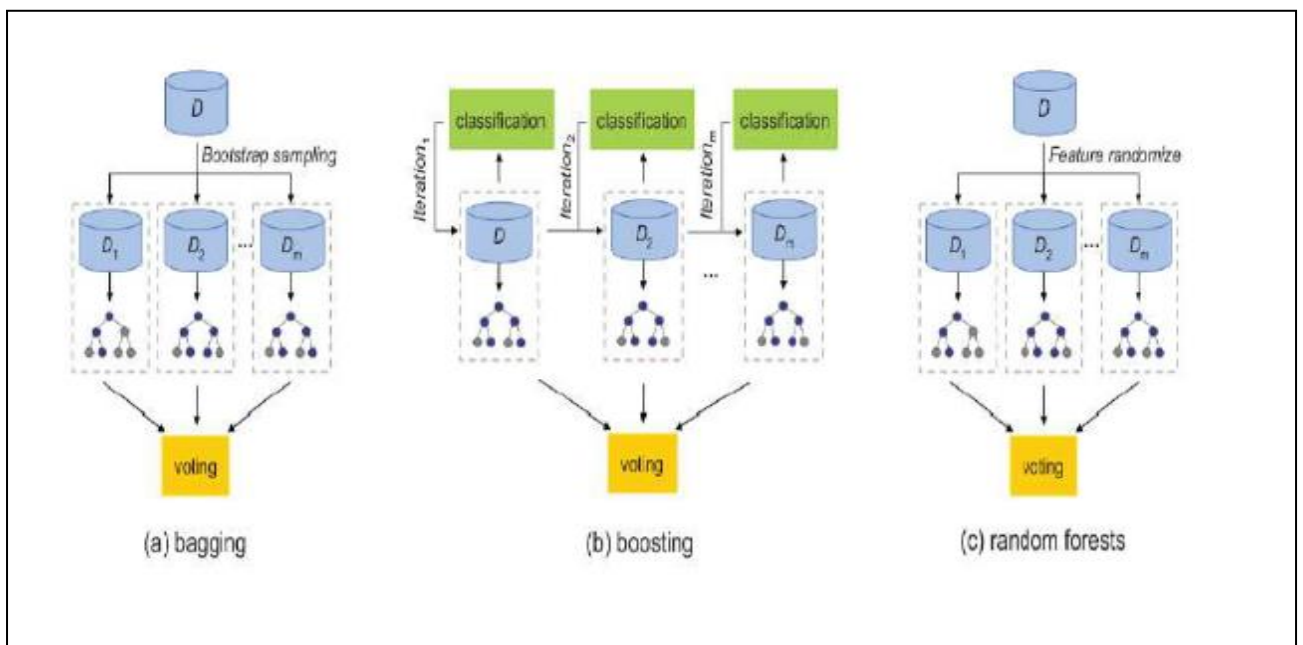


Figure II.11 : Schémas des méthodes d'ensembles de base.

II.9.Discussion

Dans ce chapitre, nous avons passé en revue quelques algorithmes basés sur l'apprentissage automatique (Machine Learning), en présentant le concept de l'apprentissage automatiques. Pour mieux élaborer notre stratégie, nous avons donné les différents types d'apprentissage automatiques et les critères adoptés pour choisir un modèle à utiliser. Pour tenter d'optimiser le fonctionnement de l'apprentissage automatique, les réseaux de neurones sont utilisés dans le deep learning en modifiant leurs architectures.

Aussi, comme nous l'avons présenté dans ce chapitre, l'apprentissage automatique a connu une amélioration en terme de performance en introduisant les méthodes ensemblistes, tels que le Bagging et le Boosting. Vu leurs performances dans la classification d'images, ces techniques basées sur l'apprentissage automatique sont très largement utilisées dans l'estimation des précipitations dont les données issues de la télédétection. Cette combinaison avantageuse, décrite dans ce chapitre, entre l'apprentissage automatique et les méthodes ensemblistes fera l'objet de notre travail. Il s'agit d'appliquer l'algorithme d'AdaBoost sur le modèle SVM en utilisant plusieurs variantes multiclassées.

Afin de mettre en évidence l'apport de ces techniques, particulièrement celles basées sur l'apprentissage automatique, sur les estimations de précipitations, dans le prochain chapitre, nous présentons une liste de travaux réalisés utilisant les données satellites météorologiques.

CHAPITRE III

Méthodes d'estimation des précipitations par satellite

III.1. Préambule :

Les informations sur les taux de précipitations, leurs quantités et leurs distributions sont primordiales dans la gestion des ressources hydrologiques et planification d'activité agricole et autres domaines dépendants. A cet effet, l'estimation de ces précipitations dans le temps et dans l'espace est d'une importance capitale. Cependant, vu le caractère aléatoire des précipitations, leur estimation est très complexe. Les moyens traditionnels utilisés pour la mesure des précipitations sont insuffisants. En effet, les pluviomètres et les radars qui fournissent des mesures directes montrent des limitations en termes de couverture.

Ainsi, l'utilisation des données satellitaires pour l'estimation et la classification des précipitations, constitue une solution très efficace pour remédier au manque de couverture des instruments de mesures au sol. En effet, de vastes régions restent en dehors de la résolution spatiale de ces instruments. Les satellites météorologiques permettent de donner un champ d'observation très étendu. Toutefois, les données issues de ces satellites ne sont pas des mesures directes des taux de précipitation. En revanche, ces données fournissent des informations sur les propriétés optiques et microphysiques des nuages, telles que la température de brillance des sommets, la phase thermodynamique, l'épaisseur et le contenu en eau et glace du nuage. Par conséquent, plusieurs méthodes ont été mises en œuvre pour tenter établir une relation entre ces propriétés optiques et microphysiques des nuages et les taux de précipitations [e.g. Adler et Negri 1988; Amorati et al.2000; Thies et al. 2008a; 2008b; 2008c ; Feidas et Giannakos 2011; 2010; Feidas et al. 2008; Levizzani et al. 2001; Lazri et al. 2013 , Kuhnleine et al 2014, Lazri et Ameer 2018].

Dans ce chapitre, nous présentons une liste de méthodes d'estimation des précipitations utilisant les données satellitaires. Pour ce faire, pour une meilleure description de ces méthodes, nous avons adopté une classification selon le type de données spectrales

utilisées ainsi qu'une séparation entre les modèles paramétriques et les modèles de machine learning.

III.2. Méthodes d'estimation des précipitations par satellite

Les méthodes d'estimation des précipitations par satellite sont des méthodes indirectes, car il n'y a pas de relations directes entre le taux de pluie et les paramètres mesurés par les capteurs de ces satellites. Elles utilisent des techniques de modélisation des liens, entre les différents paramètres fournis ou mesurés par les différents capteurs à bord des satellites météorologiques, avec les différentes résolutions spatiotemporels et spectrales. Ces paramètres donnent des informations sur les propriétés microphysiques et optiques des nuages. Ces méthodes exploitent des liens et modèles mathématiques statistiques et récemment avec l'utilisation des modèles d'apprentissage automatique pour mieux modéliser les données.

III.2.1. Méthodes infrarouge (IR) :

Ce sont des méthodes d'estimations des précipitations basées sur les données mesurées par les canaux de l'infrarouge thermique (IR) des satellites météorologiques, qui nous donnent des informations sur les températures des sommets des nuages (CTT pour Cloud Top Temperature), elles peuvent être aussi un indicateur de leurs altitudes donc hauteur du sommet de nuage (CTH :Cloud Top Height) et potentiellement de leurs épaisseurs. Ce sont des méthodes très efficaces pour détecter et estimer les précipitations dans les systèmes convectifs, qui sont l'origine d'une partie très importante des précipitations dans les régions tropicales. Elles s'appuyaient sur les considérations que les nuages avec des sommets très froids sont associés à la présence des précipitations. En conséquence, ces méthodes ont permis de relier les températures des sommets des nuages aux taux des précipitations. Parmi ces méthodes Infrarouges utilisées pour détecter et estimer les précipitations, nous citons celles basées sur le principe qu'un nuage froid est susceptible de produire plus de précipitations, et qu'il y a une relation statistique entre la fréquence d'apparition des nuages à sommet froid et les précipitations mesurés (Arkin,1979). Par exemple, voici quelques méthodes qui utilisent ce principe :

- La méthode proposée par Carn et Lahuec (1989) exploite le nombre d'occurrences (Occ) de nuages à sommets froids ($< - 40^{\circ}\text{C}$) pour estimer les précipitations cumulées sur des périodes de 10 jours.

- La méthode TAMSAT (Tropical Applications of Meteorology using SATellite data and ground-based observations) relie aussi le nombre d'occurrences (OCC) de sommets froids au taux de précipitations mesurés, mais utilise un seuil qui varie spatialement et temporellement (Milford et Dugdale, 1989). La variation de ce seuil dépend de la climatologie de la région.
- D'autres méthodes sont fondées sur l'analyse et le suivi du cycle d'activité des nuages convectifs, nécessitant des images de satellites géostationnaires et leur évolution dans le temps. La première méthode de ce type a été développée par (Griffith et al 1978).
- La technique stratiforme convective (CST) (Adler et Negri 1988) s'appuie sur un modèle de nuage à une dimension pour relier la température au sommet des nuages aux taux de pluie et à la zone de pluie, tel que les minimums locaux de température IR sont recherchés et filtrés pour éliminer les cirrus fins et non précipitants, cette méthode permet d'estimer les précipitations en tenant compte à la fois des deux types de nuages précipitants: convectifs et stratiformes, elle exploite les données Infrarouges issues du satellite géostationnaire GEOS, elle a permis la séparation entre les composantes stratiformes et convectives dans les zones où coexistent les deux types.
- La méthode GPI (GEOS Precipitation Index) qui est basé sur le seuillage de la température au sommet des nuages (Arkin et Meisner ; 1987).
- La technique NAWT Negri-Adler-Woodley (Negri et al, 1984) qui est basée sur le seuil de température des sommets des nuages, en utilisant les informations du canal IR. Elle établit un lien entre les températures des sommets des nuages et la probabilité de pluie.

Ces méthodes d'estimation des précipitations se sont avérées très utiles et efficaces pour l'estimation des précipitations des systèmes convectifs. Cependant, dans les régions de moyennes latitudes où les précipitations convectives et stratiformes cohabitent, ces méthodes ont montré leurs limites. Les précipitations stratiformes dont le sommet est relativement chaud ne peuvent pas être identifiées dans l'infrarouge. Afin de prendre en compte cette particularité, d'autres données multispectrales issues des satellites sont incorporées permettant ainsi de fournir des informations sur les propriétés microphysiques et optiques des nuages. Ceci a conduit à élaborer des méthodes exploitant les canaux Infrarouge IR, proche infrarouge NIR et vapeur d'eau WV.

III.2.2. Méthodes multi spectrales :

Ces méthodes multispectrales peuvent être appliquées dans les zones extratropicales pour prendre en compte les deux types de précipitations. Contrairement aux méthodes infrarouges, les méthodes multi spectrales exploitent les données des différents canaux, tels que le visible VIS, l'infrarouge IR, proche infrarouge NIR et vapeur d'eau WV (ex : Thies et al 2008 ; Feidas et Giannakos 2011 ; roebelling et al 2009 ; Lazri et al 2012, 2013). Nous présentons ici une liste de méthodes ayant utilisé les paramètres multi-spectraux pour l'estimation de précipitations :

- Han et al (1999) ont développé des algorithmes pour déduire l'épaisseur optique des nuages et le rayon effectif des particules précipitantes dans l'Arctique, en utilisant des données multispectrales de AVHRR (Advanced Very High Resolution Radiometer), à l'aide d'un modèle de transfert radiatif complet.
- Inoue (1985) a utilisé le canal IR et le canal NIR du radiomètre AVHRR afin de détecter les cirrus semi-transparent. Les deux bandes (10.5-11.5 μm) et (11.5-12.5 μm) ont également été employées pour la détection de cirrus non-précipitant et de cumulus dans les basses couches de l'atmosphère.
- Parol et al. (1991) ont réalisé des estimations des rayons effectifs des particules des nuages en utilisant les canaux de la bande infrarouge du radiomètre AVHRR (Advanced Very High Resolution Radiometer) du satellite NOAA.
- Kurino (1997) a identifié les nuages précipitants convectifs et stratiformes en utilisant la température de brillance $T_{\text{IR}11}$ dans le canal IR11 et les différences de température de brillance $\Delta T_{\text{IR}11-\text{IR}12}$ et $\Delta T_{\text{IR}11-\text{IR}6.7}$ issues du satellite Japonais (GMS).
- King et al (1997) ont présenté des algorithmes pour déduire certaines propriétés optiques et thermodynamiques des couches nuageuses, en particulier l'épaisseur optique, le rayon effectif des particules et la phase des particules à partir de plusieurs longueurs d'onde du radiomètre MODIS (Moderate Resolution Imaging Spectroradiometer) du satellite TERRA.
- Wolters et al. (2008) ont exploré trois algorithmes de détermination de la phase des nuages, à partir d'imageurs satellitaires passifs, pour évaluer leur pertinence à des fins de surveillance du climat dans les zones climatiques côtières des latitudes moyennes, ces algorithmes ont utilisé les données multispectrales des satellite MSG-SEVIRI et de TERRA-MODIS, les résultats obtenus de ces méthodes sont comparés aux résultats

des mesures au sol colocalisées et synchronisées de la phase des nuages, obtenues à partir des observations radar et lidar des nuages à Cabauw, aux Pays-Bas.

Des études expérimentales ont été effectuées pour étudier l'activité du rayonnement électromagnétique dans différentes bandes en fonction des propriétés optiques et microphysiques des nuages (Nauss et al, 2005 ; Kokhanovsky et Nauss 2006; roebelling et al 2006). Les simulations ont permis d'établir le lien entre les paramètres multi-spectraux des canaux des satellites météorologiques, et les propriétés microphysiques et optiques des nuages, à savoir le rayon effectif des gouttelettes d'eau (r_e) et l'épaisseur optique des nuages (τ). La combinaison de ces deux paramètres est représentée par le contenu en eau et/en glace dans les nuages (Cloud Water Path : CWP). Le CWP est définie comme LWP (Liquid Water Path) pour les nuages d'eau et comme IWP (Ice Water Path) pour les nuages de glace. Il représente la quantité d'eau (liquide ou de glace) intégrée verticalement dans le nuage (Levizzani et al. 2001), et il dépend du diamètre des gouttes de pluie et de l'épaisseur du nuage formé par ces gouttes. Il est défini par la relation suivante selon (Thies et al., 2008a ; Nauss et Kokhanovsky 2006) :

$$CWP = \rho \cdot \frac{2}{3} r_e \cdot \tau (3.1)$$

Où : ρ ($\frac{g}{m^3}$) est la densité d'eau dans les nuages.

Ce paramètre CWP est donc directement lié à la probabilité de précipitations des nuages. En suivant le modèle conceptuel, qui stipule que les nuages précipitants doivent avoir une combinaison de gouttelettes suffisamment grosses, pour pouvoir tomber facilement contre les champs de vent ascendants, et une extension verticale suffisamment grande qui favorise la croissance des gouttelettes de précipitation, et les empêche de s'évaporer sous le fond du nuage (Lensky et Rosenfeld 2003a , b), ainsi que la présence de particules de glace dans les sommets des nuages (phase de nuage).

Par conséquent, un ensemble de méthodes explorant ce concept ont été développées. Nous présentons quelques méthodes comme suit :

- Roebeling et Holleman (2009) ont présenté et validé un nouvel algorithme, pour détecter les nuages précipitants et estimer les taux de précipitations, à partir des propriétés microphysiques des nuages, extraites des paramètres multispectraux du radiomètre SEVIRI de MSG. L'algorithme des propriétés de précipitation (PP) qu'ils

ont établi, utilise des informations sur le rayon effectif des particules et la phase thermodynamique des nuages pour détecter les nuages précipitants.

- Thies et al (2008a) ont proposé une nouvelle méthode de délimitation des précipitations pendant la journée à l'aide de données satellitaires multispectrales de MSG-SEVIRI. Cette approche n'est pas seulement applicable à la détection de précipitations principalement convectives, mais permet également la détection de précipitations stratiformes en relation avec des systèmes frontaux de latitude moyenne. La technique utilise les réflectances dans les canaux VIS0.6 et NIR1.6 pour obtenir des informations sur l'épaisseur optique et le rayon effectif des nuages, et les différences de température de brillance $\Delta T_{IR8.7-IR10.8}$ et $\Delta T_{IR10.8-IR12.0}$ pour avoir des informations sur la phase des nuages.
- Thies et al. (2008b) ont élaboré une autre méthode complémentaire à la première, pour la délimitation des précipitations pendant la nuit, à l'aide de données satellitaires multispectrales issues de MSG-SEVIRI. A cet effet, ils ont utilisé les différences de température de brillance à savoir, $\Delta T_{IR3.9-IR10.8}$, $\Delta T_{IR8.7-WV7.3}$ pour remplacer les paramètres des canaux (VIS et NIR) qui ne sont pas disponibles durant la nuit et aussi $\Delta T_{IR8.7-IR10.8}$ et $\Delta T_{IR10.8-IR12.0}$ pour fournir des informations sur la phase des nuages.
- Feidas et Giannakos (2011) ont élaboré deux schémas pour la détection des précipitations à l'aide des paramètres multispectraux de MSG-SEVIRI, les schémas proposés utilisent la température de brillance (BT) dans l'infrarouge IR10.8 ainsi que les différences de température de brillance des canaux SEVIRI(BTD) : $\Delta T_{IR10.8-IR12.0}$, $\Delta T_{IR8.7-IR10.8}$ et $\Delta T_{WV6.2-IR10.8}$ comme paramètres spectraux des nuages. Deux méthodes différentes ont été utilisées pour développer les modèles de délimitation des zones de pluie. La première est une technique de seuillage des paramètres spectraux des nuages, et la seconde est un algorithme basé sur la probabilité de pluie (PoR : Probability of Rain) pour chaque pixel des données satellitaires. Ils ont constaté que l'introduction de ces différences de température des canaux SEVIRI en tant qu'informations supplémentaires dans un modèle, permet d'améliorer la discrimination des événements de pluie par rapport aux événements sans pluie par rapport au modèle BT10 à infrarouge unique.
- Lazri et al.(2013) ont utilisé le model conceptuel, basé sur le lien entre les paramètres multispectraux de MSG-SEVIRI et les propriétés microphysiques et optique des nuages, pour séparer les nuages précipitants convectifs et stratiformes dans le nord de l'Algérie. Cette technique est basée sur l'utilisation du seuillage des paramètres

multispectraux dans leur méthode CS-RADT (Convective/Stratiform Rain Area Delineation Technique).

- Lazri et Ameur (2016) ont proposé une méthode qui vise à estimer les précipitations à partir des données multispectrales de MSG-SEVIRI en utilisant le paramètre CWP (Cloud Water Path) qui représente l'épaisseur effective en eau liquide de la couche nuageuse, et la température des sommets des nuages (CTT :Cloud Top Temperature) qui sont des données extraites à partir des paramètres multispectraux de MSG-SEVIRI, qui sont les canaux Infraouge (IR) et NIR. Le principe de cette méthode consiste à étudier les relations du taux de pluie avec chacun des paramètres (CWP) et la température du sommet des nuages (CTT). Les taux de pluie sont comparés au CWP correspondant, puis aux températures de sommet de nuages CTT correspondantes. L'étude a démontré une dépendance exponentielle entre les taux de pluie et CWP, pour les taux de pluie faibles et modérés (pluies stratiformes). Inversement, les taux de pluie sont plus étroitement liés au CTT pour les forts taux de pluie (convective précipitations).

Cependant, avec l'augmentation des quantités de données et avec des résolutions spectrales et spatiotemporelles très importantes apportées par le satellite MSG, ces méthodes n'exploitent pas toute la gamme spectrale. Par conséquent, afin de mieux utiliser toutes ces informations provenant des satellites, des algorithmes basés sur l'apprentissage automatique ont été testés.

Nous introduisons dans la prochaine partie ces méthodes de Machine Learning appliquées pour l'estimation des précipitations.

III.2.3. Méthodes basées sur l'apprentissage automatique :

Les méthodes d'apprentissage automatique ont été introduites ces dernières années en raison de leurs efficacités dans la classification. Elles offrent la possibilité de traiter efficacement de grandes quantités de données et d'extraire des motifs complexes dont la relation est non linéaire. Tout d'abord, ces méthodes sont apprises en utilisant une base de données assez riche. Une fois apprises, elles peuvent être appliquées pour la séparation en plusieurs classes des données. Ainsi, elles ont permis de fournir des résultats très intéressants grâce à ce concept d'apprentissage automatique. Ces méthodes sont appliquées aux données issues des satellites météorologiques afin d'identifier les paramètres climatiques. Dans cette

partie, nous nous intéressons à leurs applications pour l'estimation des précipitations. Pour ce faire, nous nous appuyons sur la présentation de quelques méthodes les plus connues et employées dans la littérature.

III.2.3.1. Réseaux de neurones artificiels

Les réseaux de neurones artificiels (RNA) sont des modèles informatiques inspirés du fonctionnement du cerveau humain. Introduits vers les années 1990, les RNAs étaient employés dès les premières tentatives d'utilisation des méthodes d'apprentissage automatique pour estimer les précipitations à partir de données satellites, où ils ont été largement utilisés pour établir de relations entre les variables météorologiques et les propriétés des nuages précipitants. Nous présentons quelques travaux ayant utilisé les réseaux de neurones pour estimer les précipitations :

- Le projet PERSIANN (Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks) a démontré la capacité des Réseaux de Neurones à estimer les précipitations journalières avec une résolution de $(1^0 \times 1^0)$ à partir des données des satellites GOES-IR/TRMM-TMI (Sorooshian et al., 2000).
- Tapiador et al (2004) ont proposé une méthode pour estimer les précipitations à haute résolution spatiale et temporelle, en utilisant des données satellites passives. Ils ont utilisé des données de précipitations mesurées par un radar, pour entraîner un réseau de neurones artificiels, afin de produire des prévisions de précipitations à partir de données satellites passives de micro-ondes (PMW) et d'infrarouge.
- Lazri et al (2014a) ont développé une technique basée sur un réseau de neurones artificiels pour estimer les précipitations instantanées, en utilisant la température de brillance des canaux IR du radiomètre SEVIRI. Ils ont développé deux réseaux de neurones artificiels pour identifier les pixels pluvieux ou non pluvieux, ensuite pour l'attribution des taux de pluie des pixels classifiés comme pluvieux dans le premier réseau.
- Mohia et al (2017) ont élaboré une technique de classification des intensités de précipitations, basée sur les Réseaux de Neurones Artificiels, en utilisant les

caractéristiques spectrales et texturales des données satellitaires MSG. L'étude a été menée sur le nord de l'Algérie, et la méthode développée a été basée sur le réseau de perceptrons multicouches (Multi Layer Perceptron MLP). Deux MLPs ont été utilisés : le MLP-S basé uniquement sur des paramètres spectraux de MSG-SEVIRI et le MLP-ST qui utilise à la fois des caractéristiques spectrales et texturales des pixels classifiés, introduits à l'entrée de ces réseaux de neurones.

III.2.3.2. Arbre de décision

L'arbre de décision (decision trees), sont un algorithme d'apprentissage automatique supervisé, qui est construit en effectuant des séparations répétées sur les données d'entrée, en choisissant à chaque étape la variable qui permet la meilleure séparation selon un critère de mesure. Les arbres de décision sont souvent utilisés dans des problèmes de classification ou de régression.

Dans le cadre d'estimation des précipitations, les arbres de décision sont souvent utilisés pour classer les pixels des images satellitaires en fonction de leur probabilité de précipitation. Les arbres de décision peuvent également être utilisés, pour sélectionner les variables satellitaires multispectrales les plus importantes, pour la prédiction de la précipitation, en identifiant les variables qui ont le plus grand impact sur la séparation des régions de type de précipitations homogènes (feature selection). Parmi les travaux qui ont utilisé les arbres de décision, nous citons SLIQ (Supervised Learning In Quest).

- Prasad et al (2013) ont utilisé une méthode basée sur les arbres de décisions nommée SLIQ (Supervised Learning In Quest). Son principe est de procéder à la fouille de donnée basé sur les arbres de décision. La méthode est appliquée pour la prédiction et estimation des précipitations en Inde en utilisant une base de données issue de plusieurs capteurs de télédétection.

III.2.3.3. Méthode SVM (Support Vector Machines) :

Le classifieur SVM est un algorithme de machine Learning supervisé, qui peut être utilisé pour la classification ou l'estimation des précipitations à partir des données de satellites météorologiques. Comme nous l'avons décrit précédemment, les SVM sont capables de traiter des données non-linéaires et séparer les classes. Plusieurs travaux pour l'estimation des

précipitations basés sur le modèle SVM sont réalisés. Nous donnons quelques exemples ci-dessous.

- Ortiz-Garcia et al (2014) ont présenté une étude, sur les performances de la machine à vecteurs de support (SVM) dans un problème de la prévision des précipitations quotidiennes. Ils ont introduit un large ensemble de variables prédictives, telles que des données de sondage en altitude et variables dérivées d'un modèle numérique de prévision météorologique. En réalisant des comparaisons, ils ont trouvé que le modèle SVM opère efficacement comparé aux autres approches plus classiques, comme les perceptrons ou réseaux de neurones et arbres de décision.
- Sehad et al (2017) ont présenté une nouvelle technique d'estimation des précipitations appliquée au Nord de l'Algérie sur des données multispectrale de haute résolution spatiale et temporelle provenant de MSG. Ils ont proposé un schéma basé sur deux algorithmes de machine à vecteurs de support (SVM) multiclassés : un SVM_D pour les estimations de précipitations diurnes et SVM_N pour les précipitations nocturnes. Les deux modèles SVM sont entraînés à l'aide de paramètres de précipitations issues de SEVIRI, basés sur les propriétés optiques, microphysiques et texturales des nuages, pour donner les taux de précipitations journalières.
- Tebbi et Haddad 2016 ont utilisé un algorithme SVM bi-classe, avec des paramètres multispectraux de SEVIRI comme variables d'entrées, pour la classification des nuages précipitants comme convectifs et stratiformes dans le nord de l'Algérie. Les résultats obtenus sont comparés aux résultats donnés par la méthode ANN-MLP (Lazri et al 2014). Le modèle SVM a montré des performances supérieures par rapport à ANN-MLP.

III.2.3.4. Méthodes d'ensembles

Les méthodes d'ensembles, tels que l'algorithme de Bagging et l'algorithme de Boosting, utilisent une combinaison de plusieurs modèles d'apprentissage automatique de base comme apprenant faible, pour en obtenir un modèle combiné plus fort. Ils ont été largement introduits dans l'estimation des précipitations en utilisant des données de satellites météorologiques. Nous présentons une liste de travaux exploitant les méthodes ensemblistes pour effectuer des estimations des précipitations :

- Kuehnleine et al (2014 a, b) ont utilisé un classifieur RF (Random Forest) combinant plusieurs arbres de décision pour générer une réponse finale par vote majoritaire. Ces travaux ont exploré les paramètres multispectraux de MSG-SEVIRI, et les propriétés microphysiques et optiques des nuages, pour l'estimation des précipitations et la classification des nuages précipitants. Appliqué à une région de l'Allemagne situant dans les latitudes moyennes, le modèle a permis l'attribution des taux de pluie sur l'étendue des 24h.
- Ouallouche et al (2018) ont élaboré une méthode d'estimation des précipitations à l'échelle de 3 heures et l'échelle de 24 heures, au nord de l'Algérie en utilisant la méthode Random Forest (RF). Dans cette étude, ils ont utilisé les paramètres multi-spectraux de MSG-SEVIRI. La méthodologie présentée consiste en une classification des nuages en trois classes (non précipitants, convective et stratiforme) en employant RF classification. Ensuite, les auteurs procèdent par Random Forest Regression, à l'attribution des taux de pluie pour les pixels classifiés convective ou stratiformes.
- Turini et al (2021) ont mis au point une méthode basée sur les forêts aléatoires (RF) pour estimer les précipitations en Equateur. L'algorithme s'appuie sur des informations provenant de l'Integrated Multi-Satellite Retrieval for the Global Precipitation Measurement (GPM) (IMERG), et sur des données infrarouges (IR) issues du Geostationary Operational Environmental Satellite-16 (GOES-16).
- (Belghit et al 2023) Ont implémenté l'algorithme AdaBoost pour optimiser les résultats de classifications des intensités de précipitations effectuées par la stratégie One-versus-All utilisant Support Vector Machine (OvA-SVM) à l'aide des données satellite MSG. Le modèle développé qui combine l'algorithme AdaBoost avec un SVM multiclasse est appliqué après avoir testé quatre variantes différentes de SVM multiclassées avec AdaBoost, telles que l'OvO-SVM (One versus One SVM), SBT-SVM (Slant Binary Tree SVM) et DDAG- SVM (Decision Directed Acyclic Graph) aux côtés de l'OvA-SVM. L'étude a montré que l'algorithme AdaBoost fonctionnait mieux dans le cas de la variante OvA-SVM par rapport aux autres variantes. Afin d'évaluer le modèle élaboré, certaines techniques de classification, telles que la technique Enhanced Convective Stratiform (ECST), la SART où sont combinés le SVM, le réseau neuronal artificiel et les forêt aléatoire, la technique Convective/Stratiform Rain Area Delineation Technique (CS-RADT) et la technique Random Forest (RFT) sont appliquées. Les résultats de classification ont montrés que AdaBoost avec OvA-SVM (AdaOvA-SVM) a donné des performances très

intéressantes, surpassant les techniques CS-RADT, ECST et RFT. Quant à la comparaison avec le SART, la technique Ada-OvA-SVM a présenté des résultats très proches.

III.2.3.5.Méthode KNN (K nearest neighbours) :

La méthode KNN ou sa variante WKNN basée sur l'apprentissage automatique est très utilisée dans la classification. Pour prédire la classe d'un nouveau pixel, il est comparé à tous les pixels du voisinage, la méthode kNN affectera ce nouveau pixel à la classe majoritaire de ses 'k' pixels plus proches voisins, en utilisant une métrique de distance. Les performances de la méthode kNN sont dépendantes du choix du paramètre k, qu'est le paramètre principal de la méthodologie. Dans le cas des estimations des précipitations, nous énumérons ci-dessous quelques travaux ayant utilisé le KNN ou WKNN.

- Takbiri et al 2019 ont examiné l'algorithme KNN pour la détection des précipitations et de leur phase à l'aide d'observations micro-ondes passives du satellite Global Precipitation Measurement (GPM).
- Bensafi et al 2019 ont proposé une nouvelle méthode pour estimer les précipitations dans le nord de l'Algérie, qu'est basée sur l'algorithme de classification des K plus proches voisins pondérés (WkNN). Des données multispectrales du radiomètre-imageur SEVIRI de MSG sont utilisées pour déterminer le taux de pluie pour chaque pixel parmi les 16 niveaux d'intensité prédéfinis préalablement, observés dans le radar météorologique de Sétif. La classification s'est faite en utilisant les caractéristiques spectrales d'un nouvel échantillon (pixel), comme variables d'entrée du classifieur WkNN (Weighted KNN), pour prédire sa classe d'appartenance en fonction des distances pondérées le séparant des échantillons de l'ensemble des données d'apprentissage.

III.2.3.6.Méthodes à multi-classifieurs

Ces méthodes à multi-classifieurs combinent plusieurs modèles différents opérant individuellement. Les résultats sont ensuite soit fusionnés, soit combinés en utilisant des règles bien définies. Il peut y avoir différents schémas de combinaison ou de fusion de ces modèles, telle que les combinaisons en cascade successive ou les combinaisons qui sont disposées en parallèle. Ces stratégies connaissent de plus en plus de popularité dans ce

domaine de l'estimation des précipitations et des paramètres météorologiques par satellites. Dans ce qui suit, nous allons citer brièvement quelques travaux qui ont été réalisés dans le domaine de l'estimation des précipitations en utilisant les données satellitaires.

- Lazri et Aneur (2018) ont élaboré un modèle combinant trois classifieurs, à savoir Machine à vecteurs de support(SVM), Réseaux de neurones artificiels et les forêts Aléatoire. Ce modèle est conçu pour améliorer la classification des précipitations convectives et stratiformes au Nord de l'Algérie. Le modèle a été appreni puis testé sur un jeu de données multi-spectrales dérivé de MSG-SEVIRI. Les pixels bien classifiés sont les pixels ayant été attribués à la même classe par les trois classifieurs; ils sont ainsi considérés comme des pixels fiables. Tandis que les pixels moyennement ou mal classifiés sont considérés comme des pixels non fiables ayant été attribués à des classes différentes par les trois classifieurs. Pour corriger ces classifications, les pixels ont été reclassifiés à l'aide d'un nouvel apprentissage du schéma développé.
- Lazri et al (2020) ont utilisé un modèle multi-classifieur basé sur l'apprentissage automatique. Pour ce faire, ils ont d'abord combiné six classifieurs afin d'exploiter tout le potentiel de chacun d'eux. Il s'agit de Random Forest (RF1), Artificial Neural Network (ANN), Support Vector Machine (SVM), Naive Bayesian (NB), Weighted k-Nearest Neighbors (WkNN) et l'algorithme Kmeans ++ (Kmeans). L'apprentissage et la validation du modèle multi-classe, ont été effectués à l'aide des correspondances entre les données satellitaires MSG et les données radar. Six coefficients de certitude correspondant aux six classifieurs respectivement sont calculés. Dans un second temps, une autre classification a été effectuée à l'aide du classifieur Random Forest (RF2), en prenant comme paramètres d'entrée les coefficients de certitude calculés en première étape. Pour estimer ces précipitations, une régression linéaire multiple entre les classifications et les données des pluviomètres a été réalisée permettant de calculer un taux de précipitations pour chaque classe.
- Mohia et al (2023) ont mis en œuvre une méthode formée d'une combinaison de trois modèles de régression basés sur l'apprentissage automatique, en l'occurrence K-Nearest Neighbors Regression (K-NNR), Support Vector Regression (SVR) et Random Forest Regression (RFR). Pour estimer les précipitations à partir des données de télédétection, les données diurnes et nocturnes d'un pluviomètre ont été utilisées

pour l'apprentissage et la validation du modèle. Pour optimiser les résultats de leurs méthodes, ils ont combiné les sorties des trois modèles en utilisant la moyenne pondérée. Ils ont montré que la combinaison des trois modèles a nettement amélioré les prédictions.

- Belmahdi et al (2023) ont proposé une technique basée sur la théorie de Dempster-Shafer (DST) pour combiner trois classifieurs, à savoir le réseau de neurones artificiels (RNA), la machine à vecteurs de support (SVM) et la forêt aléatoire (RF) afin d'améliorer l'estimation des précipitations en utilisant les données de MSG-SEVIRI. Premièrement, les trois classifieurs ont été utilisés pour la classification des scènes de précipitations instantanées en trois classes (classe stratiforme, classe convective ou classe sans pluie). Deuxièmement, pour appliquer le DST et calculer les fonctions de masse, des probabilités normalisées à partir des classifications sont déterminées. Les fonctions de masse des différents classificateurs ont été combinées à l'aide de la DST. La méthode développée dans ce travail a montré des performances intéressantes par rapport à l'utilisation séparée des trois classifieurs.

En utilisant ces méthodes de Machine Learning, les chercheurs ont réussi à améliorer significativement les performances de l'estimation et de la classification des précipitations à partir des données satellites météorologiques par rapport aux méthodes précédentes qui étaient basées sur des approches empiriques et statistiques. Ces modèles adaptatifs s'ajustent aux caractéristiques spécifiques des données des satellites météorologiques. Elles ont donc surmonté certains défis liés à la complexité des données satellites, et leurs variabilité et diversité spatiotemporelle et spectrale, ainsi qu'à mieux modéliser les relations entre les propriétés des nuages et les taux de précipitations. Ces méthodes peuvent offrir une capacité à traiter des volumes de données de télédétection plus importants, plus diverse et variées, en temps réel et en tirer plus d'informations utiles.

Les méthodes ensemblistes exploitent l'avantage d'utilisation de plusieurs classifieurs considérés faibles. Il s'avère que cette stratégie a permis d'améliorer les performances du modèle en appliquant une série de classifications séquentielle ou parallèle permettant de se corriger mutuellement.

III.2.4. Méthodes apprentissage profond

L'apprentissage profond, est une architecture en couches sophistiquée basée sur des réseaux de neurones artificiels multicouches (LeCun et al, 2015). Il a montré son efficacité dans la reconnaissance d'images ou d'objets. Certains travaux ont été publiés utilisant le deep learning dans l'estimation des précipitations. Nous décrivons donc quelques approches basées sur le deep learning pour estimer les précipitations à partir des données satellites.

- Moraux et al (2019) ont proposé un modèle d'apprentissage profond multimodal et multitâche, pour l'estimation instantanée du taux de précipitation. Ce modèle proposé, qui est un réseau neuronal convolutif (CNN) codeur-décodeur, effectue une analyse multi-échelle de ces deux modalités qui sont : les données satellitaires thermiques IR des canaux MSG-SEVIRI (8.7, 10.8 et 12.0 μm) et les mesures automatiques de pluviomètres. Pour estimer simultanément la probabilité de pluie et la valeur du taux de précipitation pour la région couvrant l'Allemagne, les Pays-Bas et la Belgique, le modèle fusionne les différents types de données des canaux infrarouges des satellites et des pluviomètres automatiques.
- Kanchan et Shardoor (2021) ont réalisé une étude comparative de trois techniques d'apprentissage profond, pour la prévision mensuelle des précipitations dans la subdivision du Karnataka. Les performances de prévision de ces trois techniques ont été évaluées. Ces techniques sont des méthodes d'apprentissage profond telles que le réseau neuronal artificiel à propagation avant Feed Forward (ANN), le réseau neuronal récurrent simple (RNN) et la technique RNN optimisée pour la mémoire à long terme et à court terme (the Long Short-Term Memory LSTM). Les résultats et l'évaluation des performances des trois méthodes ont montré que le modèle LSTM a donné de meilleures performances par rapport à ANN et RNN.
- Tsay et al (2020) ont incorporé la méthode d'apprentissage profond « auto-encodeur convolutif » (CAE Convolutional AutoEncoder) dans la version Convolution Neural Network(CNN) de l'estimation des précipitations PERSIANN, en utilisant les données du nouveau satellite météorologique 'Himawari 8' possédant trois fois plus de canaux que ses prédécesseurs. En appliquant la technique CAE, la dérivation présentée de PERSIANN exploitée au Bureau central météorologique de Taiwan (appelée PERSIANN-CWB) étend quatre couches de convolution supplémentaires pour utiliser pleinement les canaux infrarouges et de vapeur d'eau de Himawari 8. La méthode

PERSIANN-CWB a été entraînée et calibrée à Taiwan, pour ses divers systèmes météorologiques et ses caractéristiques de précipitations localisées.

- Sadeghi et al (2019) ont exploré dans leurs étude l'efficacité de l'application de réseaux neuronaux convolutifs (CNN) avec les données des canaux infrarouges (IR) et de vapeur d'eau (WV) provenant de satellites géostationnaires GOES-West pour estimer le taux de précipitations. Les méthodes d'estimation des précipitations utilisant des réseaux de neurones artificiels (PERSIANN) – Cloud Classification System (CCS), qui est un produit opérationnel basé sur satellite, et PERSIANN – Stacked Denoising Autoencoder (PERSIANN-SDAE) sont utilisés comme modèles de base pour la comparaison. Le modèle proposé (PERSIANN-CNN), construit selon le modèle PERSIANN, fournit des résultats d'estimations plus précis que ceux des autres modèles.
- Tian et al (2020) ont appliqué des réseaux de neurones à rétro-propagation (BPNN : Back Propagation Neural Networks) et des réseaux de neurones convolutifs (CNN), pour améliorer la précision de l'estimation des précipitations à l'aide de données radar au sol. Cette technique est employée afin de résoudre le problème de la précision de l'estimation des précipitations, en utilisant la relation Z-R entre les réflectivités radar et les taux de précipitations, qui est une méthode traditionnelle pour estimation. Pour comparer les performances des méthodes Deep Learning avec le modèle de base (relation Z-R) et obtenir une meilleure solution, ils ont utilisé des données contenant deux caractéristiques principales ; les valeurs de réflectivité Radar dBZ comme entrées des modèles, et les taux précipitations correspondantes provenant des pluviomètres, comme sorties. Ils ont constaté que la méthode proposée avec CNN atteint de meilleures performances que celle utilisant BPNN et la relation Z-R.

Le deep learning est très efficace dans la reconnaissance d'images. Ces dernières années, il est massivement utilisé pour détecter des objets dans les images. Cependant, l'introduction d'apprentissage profond dans l'estimation et classification des précipitations, n'est pas totalement explorée. L'implémentation de tel concept nécessite une quantité importante de données notamment pour le Deep-ANN. Dans le cas de CNN, le traitement se fait sur une image complète ou sur un ensemble de pixels. Il ne fournit pas un traitement basé sur un pixel afin de lui attribuer une classe. Les estimations des précipitations requièrent des

résolutions spatiales importantes, et souvent à l'échelle du pixel. Par conséquent, le CNN semble être limité pour générer des estimations avec de hautes résolutions.

III.3. Méthodes Micro-ondes

Des méthodes utilisant les données micro-ondes (M-O) issues de satellites à orbite polaire pour l'estimation des précipitations sont aussi développées (Smith et al 1998 ; Wilheit et al 1994). Elles exploitent les propriétés et liens directs de ces données avec les nuages précipitants, permettant ainsi d'offrir des informations supplémentaires par rapport aux propriétés microphysiques et optiques des nuages. Elles se basent sur les principes des mécanismes physiques comme l'émission, la diffusion et la dépolarisation du rayonnement électromagnétique, due à la présence des hydrométéores dans les nuages. Ces méthodes M-O utilisent les interactions entre le rayonnement MO et les hydrométéores de la surface sondée, pour estimer son taux de précipitations, en tenant compte de la fréquence des Micro-ondes utilisées, et des propriétés de la polarisation et de l'émissivité de la surface sondée.

Les rayons Micro-ondes obtenues par les capteurs utilisés par ces méthodes directes (M-O), définissent la nature de la surface avec laquelle s'est faite l'interaction, ainsi que son état ou contenue en eau (hydrométéores) et particules précipitantes.

La fréquence d'onde envoyée, les propriétés de diffusion et d'absorption/émission des hydrométéores influencent l'interaction entre les rayons micro-ondes et les hydrométéores. Pour cela, Spencer et al (1989) ont calculé ces propriétés dans trois fréquences d'onde principales (19, 36.5 et 86 GHz), qui ont été utilisées dans ces méthodes MO pour estimer les précipitations. Ils ont remarqué que les gouttes liquides absorbent et diffusent les MO, mais l'absorption est dominante aux basses fréquences (19 et 36,5 GHz), tandis que les particules de glace n'absorbent pas le rayonnement MO quel que soit sa fréquence, mais elles diffusent un rayonnement MO important aux fréquences élevées (86 GHz). Les phénomènes de diffusion et d'absorption augmentent avec la fréquence des micro-ondes utilisées et le taux de pluie, et la diffusion par les particules de glace augmente beaucoup plus rapidement avec la fréquence (86 GHz) que la diffusion par liquide. Ainsi, si les ondes utilisées par la méthode M-O sont en hautes fréquences (86 GHz), ces méthodes dépendent alors du phénomène de diffusion par les cristaux de glace présents dans cette surface étudiée ou sondée, tandis que lorsque les M-O sont en basse fréquences (10 à 37 GHz), les méthodes dépendent de l'absorption/émission de l'eau liquide contenue dans les surfaces sondées, d'où les méthodes d'estimation des précipitations utilisant ces données Micro-ondes, sont basées soit sur le mode

émission de l'eau liquide, présent dans la surface sondée quand les micro-ondes sont de basses fréquences, soit sur le mode diffusion des cristaux de glaces contenues dans la surface ou nuages sondée dans le cas où les micro-onde sont de hautes fréquences.

Les algorithmes s'appuyant sur ces méthodes, diffèrent selon la nature de la surface visée, terre ou océan et aussi selon la longueur d'onde utilisée et le mécanisme de fonctionnement.

Des exemples de types de ces méthodes utilisés dans les estimations des précipitations, et leurs principes, ainsi que des techniques combinant ces données avec les données multispectrales de satellites géostationnaires sont détaillés dans ces références (Lazri,2013; Ana Maia-Besson,1996).

L'avantage de ce type de méthodes, est que l'information utilisée pour l'estimation des précipitations contenue dans les données MO, est directement liée aux nuages et leurs contenu en eau. Toutefois, ces données présentent un inconvénient en termes de résolution temporelle qui est très faible. Les capteurs abord de satellites à orbite polaire fournissant ces données à MO scrutent le même endroit de l'atmosphère uniquement deux fois par jour.

III.4. Méthodes combinant données MO/multispectrales

La combinaison des données Micro-ondes des satellites à orbite polaires et les données multispectrales des satellites géostationnaires a fait aussi l'objet de plusieurs travaux de recherche. Récemment, les méthodes à apprentissage automatique sont testées sur ces couples de données pour tenter d'améliorer les résultats de classification ou d'estimation. Nous résumons ci-dessous quelques travaux utilisant l'apprentissage automatique sur les données MO/multispectrales.

- Chen et al (2019) ont présenté une nouvelle méthode de fusion de données micro-ondes et multispectrales, basé sur l'apprentissage automatique pour améliorer les estimations de précipitations. L'architecture prototype de ce système de fusion est composée d'un modèle de perceptron multicouche (MLP) d'apprentissage profond, pour produire des estimations de précipitations, en utilisant comme entrées les données infrarouges (IR) des satellites géostationnaires et les données basées sur les micro-ondes passives (PMW) des satellites en orbite terrestre basse. Ils ont utilisé des mesures pluviométriques de haute qualité du réseau de radars au sol comme données étiquettes

cibles pour entraîner ce modèle MLP. Une étude de démonstration à l'échelle urbaine sur le métroplexe de Dallas-Fort Worth (DFW) est présentée.

- Turini et al (2019) ont élaboré un algorithme basé sur les paramètres infrarouges du radiomètre SEVIRI de MSG combiné aux données micro-ondes. Ils ont conçu des modèles de Forêts aléatoires (RF) utilisant les informations pluviométriques micro-ondes du produit Integrated Multi-Satellite Retrieval for the Global Precipitation Measurement (GPM) (IMERG) comme référence d'apprentissage, pour délimiter la zone de précipitation et attribuer des taux. La méthode a été validée par rapport aux données de pluie indépendantes micro-ondes de GPM IMERG, non utilisées pour la formation du modèle.

Pour essayer de tirer profit des avantages des deux types de données, tels que la résolution spatio-temporelle élevée des paramètres multispectraux et les informations sur les propriétés microphysiques et optiques des nuages, ces méthodes sont mises au point. Pour rappel, les données MO fournissent des informations directes sur les nuages précipitants.

Ces travaux montrent donc un avantage en combinant les deux types de données. Cependant, pour rendre ces méthodes fonctionnelles, les deux types de données doivent être disponibles. Aussi, l'un des inconvénients majeurs des données MO est la résolution temporelle très limitée et une couverture très réduite.

III.5.Discussion :

Dans ce chapitre, nous avons présenté différentes méthodes utilisées dans l'estimation des précipitations à partir des données satellites. Ces méthodes exploitent le caractère indirect de ces données, à savoir les propriétés optiques et microphysique des nuages. Comme nous l'avons montré, les premières méthodes utilisent uniquement les données Infra-Rouges (IR), qui donnaient des informations liées à la température et hauteurs des sommets de nuages, adaptées principalement à l'estimation des précipitations convectives. Cette limitation a conduit à l'exploitation des données multispectrales pour estimer aussi les précipitations stratiformes dont nous avons décrit quelques méthodes.

Nous avons aussi présenté les méthodes utilisant les données Micro-Ondes qui apportent des informations à liens direct avec les nuages. Toutefois, nous avons noté que ces méthodes présentaient un inconvénient majeur, à savoir la résolution temporelle des données Micro-Ondes des satellites à orbite polaire qui est très limitée.

Le contexte scientifique de l'utilisation de l'apprentissage automatique et apprentissage profond, dans ce domaine d'estimation et classification des précipitations par satellites a été introduit dans ce chapitre. Nous avons pu montrer le potentiel de machine learning en exploitant des grandes quantités de données fournies par la télédétection multi-capteurs à travers les quelques techniques que nous avons énumérées. Néanmoins, la mise en marche du deep learning nécessite une grande quantité d'informations pour le type DMLP (Deep MultiLayer Perceptron). Alors que le CNN ne peut pas fournir des estimations à des échelles du pixel. La résolution spatiale doit être ramenée à un ensemble de pixels.

Malgré le niveau de précision des résultats apporté par le machine learning, des améliorations peuvent être incorporées en limitant les taux d'erreurs de classification et d'estimation. Pour ce faire, comme il a été présenté dans ce chapitre, des méthodes ensemblistes, telles que l'algorithme de Boosting, l'algorithme de Bagging sont implémentées pour réduire soit la variance, soit le biais améliorant ainsi la qualité des résultats.

Dans cette optique, le chapitre suivant portera sur la présentation de concept des méthodes ensemblistes pour optimiser les résultats de classification et d'estimation des précipitations.

CHAPITRE IV

Estimation des précipitations par la méthode AdaBoost-OvA-SVM

IV.1. Préambule :

Dans ce chapitre, nous implémentons l'algorithme AdaBoost pour optimiser les classifications réalisées par le un-contre-tous de SVM (One-versus-All-SVM) appliqué pour la classification des intensités de précipitations en trois classes (classe convective, classe stratiforme et non précipitant). Le modèle conçu consiste en application de SVM en adoptant le 'one-versus-all' considéré comme un classifieur non fiable pour une meilleure combinaison avec l'AdaBoost. Ainsi, trois variantes du classifieur, à savoir OvA-SVM1, OvA-SVM2 et OvA-SVM3 spécialisés pour la reconnaissance des classes convective, classe stratiforme et classe non-précipitant respectivement sont boostées en utilisant l'algorithme AdaBoost.

Pour une décision finale, nous procédons ainsi ; dans le cas où un seul classifieur donne un résultat positif, la classe finale correspond à la classe donnée par ce classifieur. Si plusieurs classifieurs donnent un résultat positif, la classe finale correspond à la classe dont l'entrée est la plus éloignée de la frontière (hyperplan séparateur de SVM). De même, quand tous les résultats sont négatifs, on prend alors la classe pour laquelle l'entrée est la plus proche possible de la frontière.

Le modèle est appliqué sur une base de données constituée d'images multispectrales issues de satellite météorologique MSG (Météosat Second Generation). Les différents apprentissages sont effectués en comparant les données MSG aux données radar correspondantes. Pour évaluer les performances du modèle AdaOvA-SVM, nous avons réalisé des comparaisons avec d'autres méthodes de classifications. Ainsi, nous avons implémenté la variante OvO-SVM (One versus One SVM) avec et sans AdaBoost, la technique ECST (Enhanced Convective Stratiform Technique) de (Reudenbach, 2003; Reudenbach et al., 2001), la technique basée sur la combinaison de machine à vecteurs support, les réseaux de neurones artificiels et les forêts aléatoires (SART) de (Lazri et al, 2018), la technique 'Convective/Stratiform Rain Area Delineation Technique (CS-RADT)' de (Lazri et al, 2013), et la 'RandomForest technique (RFT)' (Ouallouche et al, 2018).

Ce chapitre est organisé comme suit :

Dans un premier temps, nous donnons la région d'étude et les couples de données utilisées. Ensuite, nous présentons le concept mathématique des différentes variantes du modèle SVM ainsi que l'AdaBoost. Dans un second temps, nous montrons l'implémentation de la combinaison de SVM et AdaBoost ainsi que les résultats des estimations de précipitations et leur interprétation.

IV.2.Zone d'étude et données :

IV.2.1.Zone d'étude :

La région d'étude est située au Nord-Est de l'Algérie (Voir figure IV.1), où elle s'étend entre les coordonnées de (32⁰N à 37⁰N de Latitude et de 2⁰W à 8⁰E de longitude) . Cette région est connue pour son climat méditerranéen où des précipitations issues des systèmes convectifs et des systèmes stratiformes cohabitent. Le climat méditerranéen connaît des situations tempérées, chaudes ou subtropicales. En été, les précipitations sont très rares voire inexistantes. L'anticyclone subtropical remonte en latitude et génère des précipitations de types convectifs. En revanche, les hivers sont bien arrosés dans cette région et connaît leur maximum de précipitations. Ces dernières sont issues en général des systèmes stratiformes.

Dans cette région, des moyens traditionnels sont déployés pour la mesure des précipitations, tels que le radar météorologique qui couvre une partie importante de la région, mais insuffisante pour concevoir des cartographies de précipitations complètes et les pluviographes qui sont fiables mais très épars, notamment pour les zones difficiles d'accès. En revanche, les données issues du satellite MSG sont collectées d'une façon régulière, continue et complète.

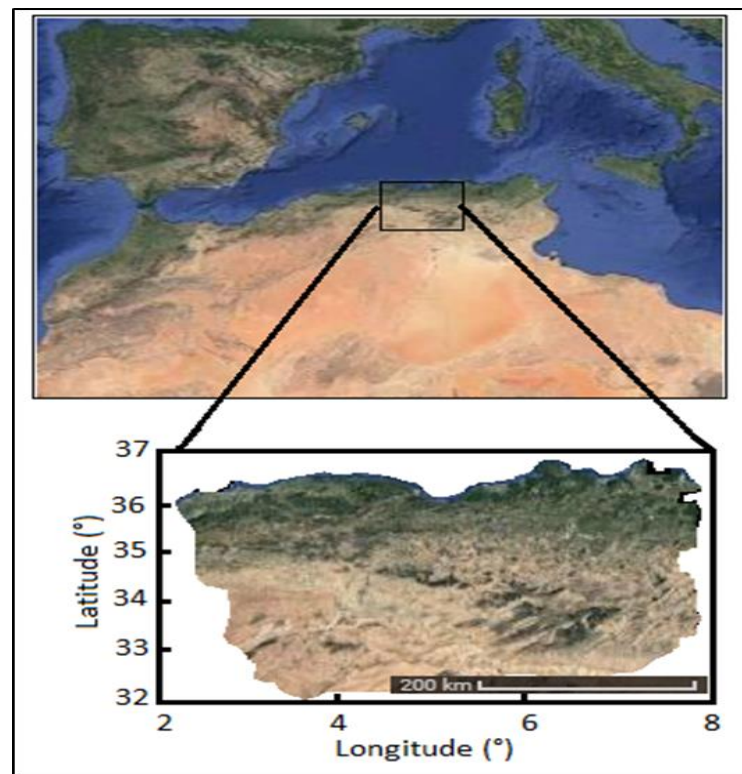


Figure IV.1. Région d'étude (Nord-Est de l'Algérie)

IV.2.2. Données :

Dans cette étude, les couples de données (données MSG/ données Radar) sont utilisées. La classification des intensités de précipitations au-dessus de la région d'étude en utilisant la combinaison AdaBoost-SVM est réalisée sur les données MSG. Quant aux données Radar, elles sont utilisées pour l'apprentissage et la validation des modèles élaborés.

Nous avons utilisé des ensembles de données de trois saisons des pluies ; pour l'apprentissage, nous avons pris toutes les données recueillies au cours des saisons des pluies (2008/2009) et (2009/2010). Pour la validation, nous avons considéré les données enregistrées au cours de la saison (2010/ 2011) (tableau IV.4). Nous avons également utilisé les données des pluviomètres pour estimer les taux de précipitations en utilisant le modèle développé. Trois échelles sont considérées, à savoir l'échelle journalière, l'échelle mensuelle et l'échelle saisonnière.

IV.2.2.1.Data MSG :

Les données MSG fournissant des informations sur les propriétés optiques et microphysiques des nuages sont implicitement utilisées pour la classification des précipitations en trois classes. L'instrument principal des satellites MSG est le radiomètre imageur SEVIRI (Spinning Enhanced Visible and Infrared Imager) qui collecte des données sur 12 canaux, une fois tous les quarts d'heure (15 minutes). Selon les canaux, la résolution spatiale de ces images est de (1 km×1km) pour le canal visible haute résolution HRV et de (3 km x 3km) pour les autres canaux aux points sub-satellite, sur la zone d'étude qui est le nord de l'Algérie, la résolution spatiale est réduite à environ $4 \times 5 \text{ km}^2$ sur cette zone (en raison de l'angle zénithal de visualisation par satellite MSG de SEVIRI).

Les différents canaux de MSG-SEVIRI sont deux canaux visibles (VIS 0.6, VIS 0.8), un proche infrarouge (NIR 1.6), huit (8) canaux dans l'infrarouge et la vapeur d'eau (IR 3.9, WV 6.2, WV 7.3, IR8.7, IR 9.7, IR 10.8, IR 12.0 et IR 13.4) et un canal visible haute résolution (HRV) à large bande ne couvrant qu'une partie du disque terrestre.

Chaque pixel est codé sur 10 bits appelé compte numérique (CN). Ce compte numérique peut prendre des valeurs de (0 à 1023) pour tous les canaux (Eumetsat, 2004). Cette valeur numérique peut être convertie en radiance et puis en température de brillance pour les canaux infrarouge et vapeur d'eau, et en réflectance pour les canaux visibles, grâce à une relation de conversion et aux informations stockées des données images RAW (niveau 1.5), c'est à dire les valeurs de taille 3712×3712 pixels de l'image (SEVIRI), et les coefficients de calibration utilisés avec le compte numérique des pixels. Pour plus de détail, le lecteur peut se référer à Eumetsat (2004).

IV.2.2.2.Données Radar :

L'ensemble de données radar est considéré comme des données de référence fiables pour l'apprentissage et l'évaluation du modèle développé. Les données sont collectées avec une résolution spatiale de (1 x 1Km²) et avec une résolution temporelle de 15 minutes par le radar du Sétif Rain Radar (SRR).

Le SRR effectue des mesures de précipitations en bande C à une fréquence opérationnelle de 5,6 Ghz. Le radar est installé près de la ville de Sétif à des coordonnées de (36° 11' N, 5° 25' E) et 1700m d'altitude. Il s'agit d'un radar qui a les caractéristiques

techniques suivantes : longueur d'onde (5,5 cm), la puissance crête de (250 kw), de fréquence de répétition (250 hz) et de durée d'impulsion (4µs).

Ce SRR fournit des images dans le 'Plan Position Indicator (PPI)' au format 512 x 512 pixels, chaque pixel est codé sur 4 bits donnant 16 niveaux possibles de réflectivités en dBZ [04, 12, 18, 22, 26, 30, 34, 38, 42, 46, 50, 54, 58, 62, 66, 70] qui est le paramètre physique du radar noté (Z) qui peut être aussi exprimé en (mm^6m^3), et peut être converti en intensité des précipitations R (mm/h), en utilisant la relation Z - R adaptée pour notre région d'étude (Lazri et al 2013a).

Cette relation entre le facteur de réflectivité radar (Z) et l'intensité des précipitations R , a été obtenue à partir de la calibration du radar météorologique, en utilisant un réseau de pluviomètres dense réparti sur la zone d'étude, lors d'un travail qui a été réalisé dans le cadre d'une collaboration, entre l'Office Nationale de Météorologie (ONM) d'Alger et le Laboratoire d'Analyse et de Modélisation des Phénomènes Aléatoires (LAMPA) de l'université de Tizi-Ouzou (Lazri et al 2013a).

Cette relation entre l'intensité des précipitations R (mm/h) et le facteur de réflectivité du radar (Z) est donnée dans le tableau suivant :

Type	Relation	Coefficient de Corrélation
Précipitations stratiformes $Z \leq 38$ dBZ	$Z = 248.28 \times R^{1.58}$	0.79
Précipitations convectives $Z \geq 42$ dBZ	$Z = 412.35 \times R^{1.49}$	0.84

Tableau IV.1 : Relation entre l'intensité des précipitations et la réflectivité radar (Z - R).

IV.2.2.3. Données pluviographiques :

Nous avons exploité les données d'environ 201 stations pluviométriques, qui sont réparties sur la zone d'étude. Toutes ces stations fournissent des mesures directes des précipitations à l'échelle journalière. Nous pouvons calculer les accumulations de précipitations pour une période donnée en utilisant des mesures quotidiennes. Ces données fournies par les pluviomètres sont utilisées dans les estimations des précipitations par satellite.

Pour faire correspondre ces mesures ponctuelles avec les données satellitaires, nous avons utilisés les coordonnées géographiques des pluviomètres. Afin d'avoir le maximum de correspondance, chaque pluviomètre est collocalisé avec une fenêtre de 3×3 pixels (pixels moyens) situé à l'aplomb (verticalement au-dessus) du pluviomètre.

IV.2.2.4. Correspondance entre les données satellite et les données Radar :

Pour faire correspondre les données satellites aux données radar, des prétraitements spatio-temporels sont donc réalisés. Pour la comparaison spatiale entre les données radar et satellite, les données radar avec une résolution spatiale originale de $(1 \times 1 \text{ km}^2)$ ont été reprojctées et ré-échantillonnées à la résolution spatiale des images SEVIRI dans la zone d'étude qui est d'environ $(4 \times 5 \text{ km}^2)$.

Afin de réduire le décalage de parallaxe et d'erreurs de collocalisation, pour les nuages des niveaux supérieurs, nous avons moyenné les pixels de SEVIRI sur des fenêtres de (5×5) pixels pour les nuages hauts et des fenêtres de 3×3 pixels pour les nuages des étages inférieurs.

En revanche, du fait du décalage temporel faible qu'est très minime (3 min), il n'est pas nécessaire d'effectuer une synchronisation temporelle.

IV.2.2.5.Sélection des paramètres d'entrées

Pour optimiser la classification des précipitations, dans notre zone d'étude où les deux systèmes convectifs et stratiformes coexistent, les propriétés microphysiques et optiques des nuages devront être prises en considération. En effet, les propriétés optiques et microphysiques des nuages considérées pour la caractérisation des précipitations sont :

- L'extension verticale des nuages.
- L'existence de particules de glace dans les nuages.
- La température des sommets des nuages.
- La taille des particules dans les nuages.
- L'épaisseur optique des nuages.
- L'évolution du nuage.

Pour avoir des informations sur ces propriétés optiques et microphysiques des nuages, nous avons sélectionné les différents canaux de SEVIRI suivants : *VIS0.6*, *NIR1.6*, *IR3.9*, *WV6.2*, *WV7.3*, *IR8.7*, *IR10.8* et *IR12*. En fonction de leur disponibilité, nous avons considéré des combinaisons pour le jour et d'autres pour la nuit (voir tableau IV.2). Ces combinaisons sont utilisées comme entrées de notre modèle.

Paramètres d'entrée	Valeurs Possibles		Information sur
	Jour	Nuit	
$T_{10.8}$ (Kelvin)	207.2k à 283.9k	205.3k à 282.4k	Extension verticale des nuages (CTH) et température des sommets des nuages (CTT) (Thies et al 2008a ; b, Feidas et Giannakos 2011).
$\Delta T_{10.8-12.0}$ (Kelvin)	-0.3k à 7.4k	-0.3k à 7.1k	Existence de particules de glace dans les nuages (Thies et al, 2008a ;b)
$\Delta T_{8.7-10.8}$ (Kelvin)	-4.6k à 1.3k	-4.8k à 1.7k	Existence de particules de glace dans les nuages, phase thermodynamique des nuages (CP) (Nauss et Kokhanovsky 2006; rebelling et al 2006)
$\Delta T_{6.2-10.8}$ (Kelvin)	-50.1k à 6.4 k	-51.8k à 5.1 k	Extension verticale des nuages, température des sommets des nuages (Levezanni 2003, Feidas et Giannakos 2011).
$\Delta T_{7.3-12.1}$	-50.3k à 6.6 k	-52.0k à 5.7 k	Extension verticale des nuages, température des sommets des nuages (Feidas et Giannakos 2011).
$R_{0.6}$ (μm)	0.02 μm à 1 μm	Non utilisé	Taille des particules dans les nuages et Epaisseur optique des nuages (Kuhnlien et al 2014; Thies et al, 2008a)
$R_{1.6}$ (μm)	0.03 μm à 1 μm	Non utilisé	Taille des particules dans les nuages et Epaisseur optique des nuages (Kuhnlien et al 2014 ; Thies et al, 2008a)
$\Delta T_{3.9-10.8}$ (Kelvin)	Non utilisé	-10.3k à 15.1k	Taille des particules dans les nuages et Epaisseur optique des nuages (Kuhnlien et al 2014 ; Thies et al, 2008b)
$\Delta T_{3.9-7.3}$ (Kelvin)	Non utilisé	-4.9 k à 25 k	Taille des particules dans les nuages et Epaisseur optique des nuages (Kuhnlien et al 2014 ; Thies et al, 2008b)
$\Delta T_{10.8(t)-10.8(t-1)}$	-75.8k à 75.4k	-76.4k à 76.2k	Evolution du nuage (Lazri et al, 2014 ; Oukali et al, 2019)

Table IV.2. Paramètres spectraux choisis comme entrée de notre modèle et leurs valeurs possible et les informations qu'ils fournissent.

Pour la sortie du modèle développé, nous avons pris trois classes de précipitations fixées à partir des données radar, il s'agit de la classe convective, classe stratiforme et non précipitant. Le tableau IV.3 montre les intervalles de réflectivité radar en dBZ adoptés pour les différentes classes.

Réflectivités radar	Classes
pixel>38dBZ	Convective
04dBZ<pixel<=38dBZ	Stratiforme
Pixel <=04dBZ	Non précipitant

Tableau IV.3 Classes des précipitations et intervalles correspondants en dBZ.

Comme le montre le (tableau IV.1) , la réflectivité radar de 38dBZ a été adoptée pour séparer entre les précipitations stratiformes et les précipitations convectives (Lazri et al 2013 a, 2013 b). En effet, les pixels avec une réflectivité radar égale ou inférieure à 38dBZ sont considérés comme des pixels stratiformes (classe stratiforme). En revanche, les pixels au-dessus de ce seuil sont étiquetés pixels convectifs (classe convective). Les pixels considérés comme non précipitants (classe no rain) sont ceux qui ont la valeur de réflectivité inférieure ou égale à 04dBZ.

IV. 3. Contexte scientifique de l'étude :

La méthode élaborée dans ce travail, qui est une méthode ensembliste peut être appliquée à tous les algorithmes instables, ceux pour lesquels un petit changement dans les données d'entraînement induit un grand changement dans la classification finale. Ce concept est très utilisé pour améliorer les classifications obtenues à partir d'un seul classifieur. Plusieurs apprentissages d'un classifieur selon des itérations en utilisant des ensembles de données, quand les résultats sont combinés, produisent un meilleur résultat. La classification finale est obtenue soit, on prenant une moyenne des résultats données par les différentes itérations ou par vote majoritaire des classifications individuelles (Dietterich 2000).

Des travaux réalisés par (Breiman 1998 ; 2000) et (Opitz et Maclin1999) ont mis en évidence les performances des méthodes ensemblistes en décomposant l'erreur. L'erreur de prédiction peut être décomposée en trois, à savoir le biais, le taux d'erreur minimum et la variance. Le biais permet d'indiquer à quel point la classification moyenne obtenue par l'algorithme est éloignée de la donnée réelle. Quant à la variance, elle indique les différences des résultats entre plusieurs classifieurs ou itérations d'un même algorithme, c'est-à-dire les

classifieurs de base résultant des itérations composant cette méthode d'ensemble. Donc pour améliorer les classifications, et vu qu'il n'est pas possible d'influencer le taux d'erreur minimale, les méthodes ensemblistes visent à diminuer la variance ou/et le biais.

Dans ce contexte, la façon la plus commune pour diminuer ces deux paramètres est de modifier légèrement l'ensemble d'entraînement, en introduisant un nouveau paramètre variable ou méthode pour produire des ensembles différents, qui permettent de construire des classifieurs dont les résultats seront combinés pour générer une meilleure classification. Une des méthodes la plus populaire est le Boosting. L'algorithme de Boosting a été développé par Schapire (1990), en adoptant une variante de cet algorithme qui est l'algorithme AdaBoost (Adaptive Boosting) de (Freunde et Schapire, 1997) qui est un algorithme qui combine plusieurs modèles entraînés séquentiellement d'un classificateur non fiable pour générer une décision finale robuste et un classifieur efficace. Le terme non fiable est utilisé dans le sens où le classificateur utilisé comme apprenant de base pour AdaBoost, réalise des prédictions légèrement supérieures au jeu de hasard. Il n'est donc pas nécessaire de prendre un classifieur robuste, voire n'est pas carrément recommandé. La condition pour que AdaBoost soit performant utilisé sur un classifieur, est que ce classifieur ne soit pas robuste mais assez précis qu'un classifieur aléatoire. Ce qui signifie que plus deux apprenants faibles deviennent précis, moins ils peuvent être en désaccord les uns avec les autres, donc moins le AdaBoost aurait de possibilité de bénéficier de la variance des classifieurs, et de leurs résultats variés pour en choisir le meilleur. Autrement dit, l'utilisation d'AdaBoost avec un classifieur performant comme apprenant faible, risque de fausser les résultats de prédiction ou de donner des classifications moins précises.

Dans ce sens, l'implémentation d'AdaBoost sur le classifieur SVM est controversée et débattue. Des travaux montrent l'efficacité d'AdaBoost sur SVM (Cheng et Jhan, 2013), et les expériences sur des ensembles de données déséquilibrés ont montré que l'AdaBoost-SVM fonctionnait bien mieux que le SVM sans AdaBoost (Li et al, 2005). D'autres considèrent que le SVM est un classifieur Robuste et donc ne peut être optimisé par l'AdaBoost (Wickramaratna et al, 2001).

Dans ce travail, l'un des objectifs est de répondre à cette problématique, à savoir la possibilité d'obtenir un classifieur d'ensemble efficace en utilisant AdaBoost avec SVM comme apprenant faible. Nous utilisons donc l'algorithme d'AdaBoost sur le classifieur SVM pour vérifier la faisabilité de cette combinaison (AdaBoost et SVM).

Pour ce faire, nous implémentons une certaine stratégie de classification en utilisant le SVM. Quatre variantes de SVM multi-classes sont considérées à savoir : le SVM-Un-Conte-Tous (OvA-SVM), le SVM-Un-Contre-Un (OvO-SVM), SBT-SVM et DDAG-SVM. En général, ces variantes présentant un fonctionnement différent, ne donnent pas les mêmes résultats, et l'algorithme AdaBoost sera testé sur toutes ces variantes, pour voir la possibilité d'utiliser AdaBoost avec SVM, et quelle variante de SVM multi-classe serait la mieux adaptée qui donnera les meilleurs résultats avec AdaBoost.

IV.4. Méthodologie

Pour répondre à ce double objectif, à savoir la classification des intensités de précipitation en utilisant les données de MSG, et la mise en évidence de l'apport de l'utilisation de l'algorithme AdaBoost sur l'optimisation des classifications obtenues, nous avons implémenté quatre stratégies multi-classes qui sont données comme suit :

- Un-Contre-Tous-SVM (OvA-SVM)
- Un-Contre-Un-SVM (OvO-SVM)
- Arbre Binaire Oblique SVM (Slant Binary Tree SVM (SBT-SVM))
- SVM de graphe acyclique dirigé par décision (DDAG-SVM).

Pour toutes ces variantes, l'algorithme AdaBoost est appliqué en aval pour étudier la compatibilité d'AdaBoost avec SVM. Avant l'élaboration de notre démarche, nous donnons les concepts mathématiques de l'Algorithme AdaBoost et du classificateur SVM.

IV.4.1. Concept mathématique d'AdaBoost :

AdaBoost (Adaptive Boosting) est un algorithme itératif introduit initialement par Freund et Schapire (1997). C'est est une méthode d'apprentissage d'ensemble qui tente de construire un algorithme d'apprentissage "fort" basé sur un groupe de classifieurs "faibles".

L'idée principale d'AdaBoost est d'exécuter à plusieurs reprises d'une manière itérative un algorithme d'apprentissage faible, donné dans différentes distributions de probabilité W , sur les données d'apprentissage. Cette distribution est initialement uniforme ou tous les éléments de l'ensemble de données ont le même poids. Ensuite il attribue des poids plus élevés aux échantillons mal classifiés par le classifieur faible, dans le but que le nouveau

classifieur faible de l'itération suivante puisse réduire l'erreur de classification, en se focalisant sur ces pixels mal classifiés.

Dans le même temps, des poids inférieurs seront attribués aux pixels correctement classifiés. Par la suite, la distribution W est mise à jour après chaque cycle ou itération. Au final, les hypothèses émises par l'apprenant faible de chaque itération pondérée selon la précision de chacun de ces apprenants, sont regroupées en une seule hypothèse « forte », qui sera le résultat final d'Adaboost.

Il convient de noter que le Boosting a été très efficace pour résoudre les problèmes de classification à deux classes. Pour obtenir une classification multi-classes, nous effectuons plusieurs classifications binaires.

L'AdaBoost est mathématiquement formulé comme suit :

Soient les couples $(x_1, y_1), \dots, (x_n, y_n)$ où $x_i \in X$; sont les caractéristiques de l'individu i ou ses attributs (features) et $y_i \in Y = \{-1, +1\}$ la variable à prédire ou la classe de x_i .

L'algorithme attribue initialement le poids associé à i (les poids de chaque échantillon des données à classifier) :

$$D_t(i) = 1/n, i = 1, 2, \dots, n \quad (4.1)$$

Pour : $t = 1, 2, \dots, T$:

- Déterminer la fonction (hypothèse) $h_t : X$ vers $\{-1, +1\}$, qui minimise l'erreur de classification ε_t en fonction des poids D_t des échantillons classifiés selon les équations suivantes :

$$h_t = \arg \min_{h \in H} \sum_{i=1}^n D_t(i) [y_i \neq h(x_i)] \quad (4.2)$$

$$\varepsilon_t = \sum_{i=1}^n D_t(i) [y_i \neq h(x_i)] \quad (4.3)$$

ε_t est l'erreur de l'apprenant (erreur du classifieur faible de l'itération t)

- Si $\varepsilon_{\min, t} < 0.5$ la fonction (apprenant) est sélectionnée, sinon l'algorithme s'arrête

- AdaBoost calcule alors le pas de gradient (ou le coefficient de l'apprenant h_t) : $\alpha_t \in \mathbb{R}$, avec :

$$\alpha_t = \frac{1}{2} \ln \frac{1 - \varepsilon_t}{\varepsilon_t} \quad (4.4)$$

- Il met ensuite à jour le poids des exemples pour la prochaine itération:

$$D_{t+1}(i) = \frac{D_t(i) e^{-\alpha_t y_i h_t(x_i)}}{Z_t} \quad (4.5)$$

Où $Z_t = \sum_{i=1}^N D_t(i) e^{-\alpha_t y_i h_t(x_i)}$ choisi de telle sorte $D_{t+1}(i)$ soit une distribution.

Quand l'algorithme s'arrête à l'itération K, le classifieur résultant du processus de sélection est :

$$H(x) = \text{sign} \left(\sum_{t=1}^K \alpha_t h_t(x) \right) \quad (4.6)$$

H est le résultat final de ce classifieur, obtenue par la combinaison pondérée des résultats $h_t(x)$ tous les apprenants faibles des itérations de AdaBoost.

IV.4.2. Concepts mathématiques de Support Vector Machines (SVM) :

Le modèle SVM a été développé à partir de la théorie de minimisation des risques structurels (Cortes et Vapnik, 1995 ; Crammer et Singer, 2002). Dans un problème de classification binaire, la fonction de décision de SVM est

$$f(x) = \langle w, \phi(x) \rangle + b \quad (4.7)$$

Où ϕ est une fonction qui représente les données x dans un espace de dimension supérieure à l'espace d'entrée , dans le cas où la séparation linéaire des classes de données n'est pas possible dans le plan de données d'entrée , en transformant les données dans un autre espace (éventuellement dimensionnel infini) R^D :

$$\phi : R^d \longrightarrow R^D \quad \text{avec : } D \gg d$$

D est appelé le nouvel espace de caractéristiques défini par la fonction ' ϕ ' .

Où $\phi(x)$ est est la transformée de x de l'espace d'entrée à un espace de caractéristiques de plus grande dimension. Le $\langle ., . \rangle$ désigne le produit scalaire dans le nouvel espace des caractéristiques de plus grande dimension. Les valeurs optimales de w et b peuvent être obtenues en résolvant le problème d'optimisation suivant :

$$\text{Minimiser : } g(w, \delta) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^N \delta_i \quad (4.8)$$

$$\text{Sous réserve de : } y_i (\langle w, \phi(x) \rangle + b) \geq 1 - \delta_i, \delta_i \geq 0 \quad (4.9)$$

Où δ_i est la $i^{\text{ème}}$ variable d'écart toléré ou de relâchement (slack) et C est le paramètre d'ajustement, tel que C est le paramètre qui contrôle le compromis entre l'erreur tolérée dans la classification et la séparation des échantillons δ_i , et l'élargissement de la marge.

Selon la forme dual de Wolfe, le problème de minimisation (équations 4.8 et 4.9) peut être écrit comme suit (équations 4.10 et 4.11):

$$\text{Minimiser : } W(\alpha) = - \sum_{i=1}^N \alpha_i + \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N y_i y_j \alpha_i \alpha_j k(x_i, x_j) \quad (4.10)$$

$$\text{Sous la contrainte : } \sum_{i=1}^N y_i \alpha_i = 0, \forall i: 0 \leq \alpha_i \leq C \quad (4.11)$$

Où α_i est un multiplicateur de Lagrange qui correspond à l'échantillon x_i .

$k(.,.)$: est une fonction noyau , qui mappe (transforme) implicitement les vecteurs d'entrée et le produit scalaire entre deux vecteurs x_i, x_j dans un espace de plus grande dimension de caractéristiques approprié . Cette fonction $k(.,.)$ nous donne le produit scalaire de ces vecteurs dans le nouvel espace de fonctionnalité de dimension supérieure (équation 4.12) , sans avoir à calculer les vecteurs de transfert $\phi(x_i)$ et $\phi(x_j)$ de ces vecteurs vers ce nouvel espace de caractéristiques .

$$k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle \quad (4.12)$$

SVM calcule automatiquement le nombre et l'emplacement des centres, des poids et des seuils en utilisant une fonction de noyau appropriée. Pour la classification avec SVM, en général, un noyau gaussien ou un noyau polynomial peut être utilisé.

Dans ce travail, c'est le noyau Gaussien RBF (Radial Basis Function) qu'est utilisé (figure IV.2), car il présente de meilleurs résultats, il est donnée par l'expression suivante

$$k(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \quad (4.13)$$

Les échantillons sont cartographiés (transformés) de manière non linéaire dans un espace de caractéristiques (fonctionnalités) de plus grande dimension. Dans cet espace, où un hyperplan séparateur optimal est construit par les vecteurs supports. Les vecteurs de support correspondent aux centres des noyaux RBF dans l'espace d'entrée, ils sont ces vecteurs se trouvant aux frontières limites des classes. C'est l'écart entre ces vecteurs qui constitue la marge.

Les performances de généralisation de SVM sont principalement affectées par les paramètres du noyau, par exemple σ , et le paramètre d'ajustement C . ces paramètres doivent être définis au préalable.

L'approche SVM est formellement définie pour les problèmes à deux classes. Pour le cas d'un SVM multi-classe, il suffit de l'étendre au problème à K -classes avec $Y \in \{Y_1, \dots, Y_k\}$. Il existe quatre variantes principales pour la conception d'un SVM multiclasse (Du et al, 2020).

➤ **Approche ‘one versus all’ (Un-contre-Tous) :**

Dans ce cas, il faut construire tour à tour K modèles discriminants avec la modalité y_k contre les autres $Y' \in \{Y_k=+1, \dots, Y_{(k)}=-1\}$, tel qu'on considère la classe y_k comme la classe positive et toutes les autres classes comme la classe négative.

Nous obtenons K fonctions de décision $f_k(x)$. En classification, pour chaque élément x , la classe présentant le score le plus élevé est sélectionnée en utilisant l'expression suivante :

$$\hat{y} = \arg \max_k f_k(x) \quad (4.14)$$

Où \hat{y} est la fonction qui donne la classe à laquelle est affecté l'élément x classifié par cette méthode, telle qu'elle affecte l'élément classifié à la classe qui est donnée avec la plus grande possibilité donc les résultats positifs le plus grand.

$f_k(x)$: est la fonction de décision de chacun des K modèles binaires composant cette variante de 'one-versus-all -SVM', elle donne la probabilité d'appartenance de x à la classe 'k'.

L'implémentation de cette variante OvA-SVM nécessite K apprentissages à effectuer sur les données, où à chaque apprentissage, une classe est distinguée de toutes les autres. Dans le cas où plusieurs modèles parmi les K modèles aient un résultat positif en classifiant un élément, donc pour affecter l'élément à une des K classes, on prend celui qui est le plus certain de son résultat (c'est-à-dire celui pour lequel le point est le plus éloigné de la frontière de la marge ou l'hyperplan). De même, quand tous les résultats sont négatifs, on prend alors le modèle pour lequel l'entrée est le plus près possible de la frontière (c'est-à-dire, celui pour lequel on est le moins sûr que le résultat est négatif).

Cependant, pour cette variante comme indiqué précédemment, un déséquilibre des classes dans la construction des modèles individuels conduit à un mauvais apprentissage.

Avec cette stratégie, pour notre cas à trois classes, trois classifications binaires (SVM binaires) en parallèle sont effectuées (voir Figure IV.3a) ; classification1 (convectif contre stratiforme et 'non précipitant'), classification2 ('stratiforme' contre 'convectif' et 'non précipitant'), classification3 ('non précipitant' contre 'convectif' et 'stratiforme').

➤ **Approche One versus One (Un-Contre-Un) :**

Une autre variante de SVM multiclasse est l'approche one versus one (OvO-SVM). Il s'agit dans le cas où on a K classes, de construire tour à tour ' $K(K-1)/2$ ' modèles discriminant chaque paire de classes en considérant la $K^{i\text{eme}}$ classe comme classe positive et l'autre classe comme négative ($Y' \in \{y_k=+1, y_j=-1\}$). Nous obtenons ' $K(K-1)/2$ ' fonctions $f_{k,j}(x)$, où cette fonction nous donne la probabilité que l'élément classifié x appartient à la classe K comparé à la classe J .

En classement, pour la classification finale d'un élément x la classe est prédite en utilisant un système de vote c.-à-d. présentant le plus grand nombre de victoires (équations 4.15 et 4.16), donc l'élément x sera affecté à la classe qui s'est vue affecté cet éléments x le plus grand nombre de fois comparé a toutes les autres classes j .

$$D_k(x) = \sum_{j \neq k, i=1}^k \text{sign}[f_{k,j}(x)] \quad (4.15)$$

$$\hat{y} = \arg \max_k D_k(x) \quad (4.16)$$

Dans le cas d'ex-æquo (2 classes ou plus ont le même nombre de votes), on procède à la somme des scores $f_{k,j}(x)$ et on prend le maximum. L'implémentation de cette variante nécessite $K(K-1)/2$ apprentissages ou modèles à effectuer, mais avec des ensembles de données de taille réduite car à chaque apprentissage qui compare entre deux des K classes , le modèle utilise les données d'apprentissages de ces deux classe uniquement . En général, cette variante à moins de problème de déséquilibre des classes, les modèles sont mieux calibrés.

Pour notre cas, trois classes (convective, stratiforme et non précipitant) sont l'objet de notre classification. Par conséquent, trois classifications binaires (SVM Binaire) en parallèle sont possibles (voir figure IV.3b) : classification1 (convective/stratiforme), classification2 (convective/non précipitant), classification3 (stratiforme/non précipitant). Une décision finale est produite selon l'équation (4.15).

➤ **Variante «Decision Directed Acyclic Graph (DDAG) » :**

Cette stratégie est schématisée par un graphe orienté fini, le DDAG est construit à partir de nœuds de décision de nombre $N \times (N - 1)/2$ classifications binaires pour un problème de classification de N classes, où chaque nœud excepte (exclut) une classe de toutes les autres classes, dans le cas où le nœud a choisi une classe, l'autre est éliminée.

Les nœuds dans DDAG sont organisés en triangle avec le nœud racine unique en haut et croissant, ensuite par incrémentation d'un nœud dans chaque couche, jusqu'à la dernière couche qui aura N nœuds selon le nombre de classes.

Dans cette étude, qui porte sur trois classes (convective, stratiforme, non précipitant), deux niveaux de classification séquentielle sont appliqués. Le premier niveau contient un nœud unique qui se sépare en deux classes (non convective et non stratiforme). Le deuxième niveau contient deux nœuds en parallèle, un nœud divise le 'non convective' en deux autres classes ('non stratiforme = 'non précipitant' et 'non non précipitant' =stratiforme'). L'autre nœud est appliqué à 'non stratiforme'. Il se sépare en deux classes (non convective = non precipitant et 'non non precipitant = convective'). Dans chaque nœud, le SVM binaire est implémenté. La (figure IV.3d) explique cette stratégie à trois classes.

➤ **Variante Slant Binary Tree (SBT) (Arbre Binaire Inclinée):**

Cette stratégie consiste à diviser en deux classes à chaque nœud de l'arbre pour construire une classification multi-classe à la fin des divisions. Cette stratégie peut être élaborée en deux architectures : arbre binaire incliné (SBT) et arbre binaire équilibré (Balanced Binary Tree BBT) (Du et al, 2020).

Dans le cas de cette étude, nous avons utilisé l'architecture des arbres binaires inclinés (SBT). Le nombre de classes considérées au départ est de trois classes. Le premier nœud sépare la classe 'non-précipitant' de la classe 'précipitant'. Le deuxième nœud divise la classe 'précipitant' en deux classes 'la classe convective' et 'la classe stratiforme'. Le classifieur implémenté dans ces nœuds est le SVM binaire (bi-classe) (voir figure IV.3 c).

Il n'y a pas de grande différence notée entre l'utilisation du noyau polynomial ou du noyau gaussien dans les variantes SVM, mais nous avons notés une légère amélioration de la précision de classification, pour le SVM utilisant noyau gaussien comme le montre la (Figure IV.2). Comme nous l'avons mentionné précédemment, nous avons opté pour la fonction de noyau gaussien (RBF) dans tous nos SVM utilisés dans ces variantes.

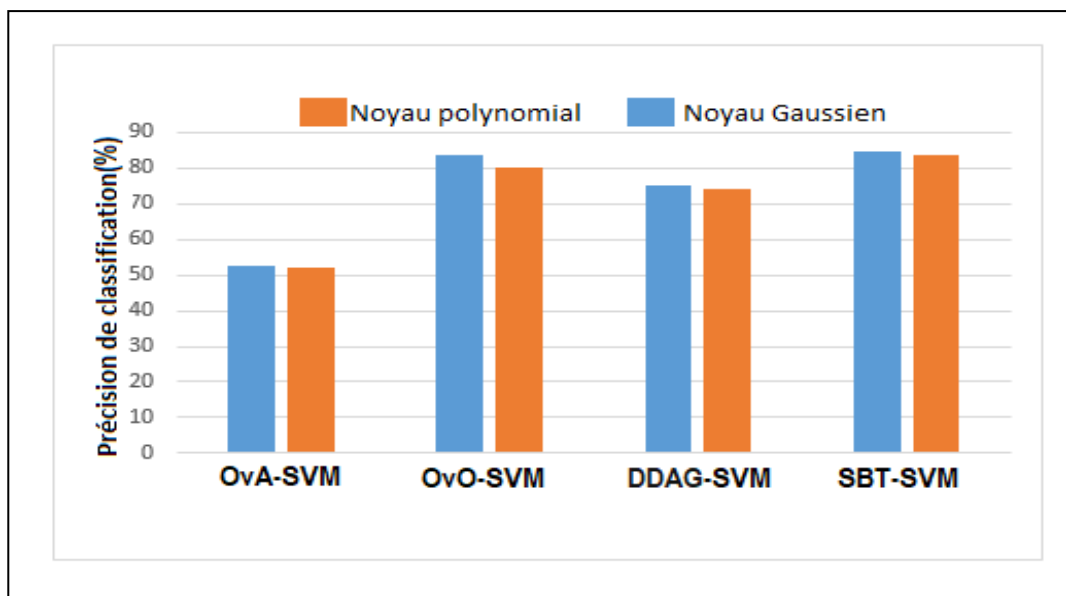


Figure IV.2 : Précision de la classification pour l'application de différentes variantes SVM multiclases utilisant le noyau polynomial par rapport au noyau gaussien dans la classification des précipitations.

La figure IV.3 montre les schémas des différentes variantes (OvA, OvO, DDAG et SBT) utilisées pour la construction de SVM multiclass.

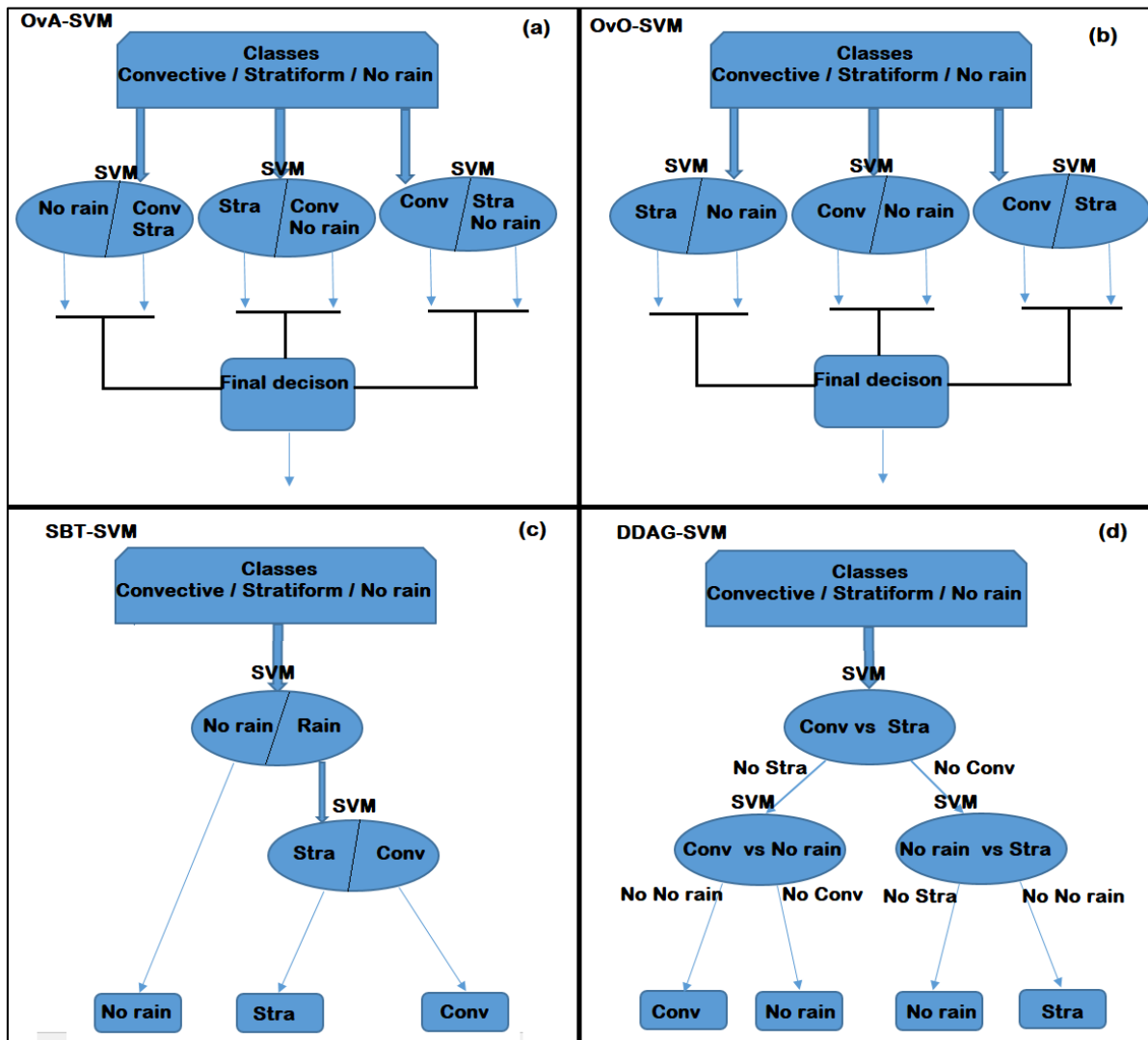


Figure IV.3 : Différentes stratégies utilisées pour la construction de SVM multiclass, a) OvA-SVM multiclass, b) OvO-SVM multiclass, c) SBT-SVM multiclass, d) DDAG-SVM multiclass.

IV.4.3. Elaboration des modèles de classification avec AdaBoost :

L'implémentation des algorithmes AdaBoost est plus efficace dans le cas des classifications binaires. Ils ont été initialement conçus pour des classifications binaires (Freund et Schapire, 1997).

Ainsi, pour mieux optimiser la classification, pour chaque séparation binaire par SVM Non-Linéaire dans nos schémas de classifieurs SVM, l'algorithme AdaBoost est appliqué. Comme mentionné dans la section IV.4.2, pour implémenter le SVM non linéaire, le noyau gaussien RBF (Radial Basis Function) est utilisé.

Pour résoudre des problèmes multi-classes, il suffit d'appliquer une série de classifications binaires, en adoptant les quatre stratégies (OvO, OvA, SBT et DDAG) pour construire le SVM multi-classes avec des SVM bi-classes pour la classification des types de précipitations à partir des images MSG (non-précipitant /convectif/ stratiforme).

Nous avons donc appliqué pour chaque sortie de classification binaire SVM une routine de l'algorithme AdaBoost pour toutes les variantes de SVM mentionnées en haut. La figure IV.4 met en évidence les différentes stratégies combinant AdaBoost et SVM.

La figure (IV.4a) montre la combinaison de l'algorithme AdaBoost avec le multi-classifieur One-versus-One SVM. Il s'agit de l'application de l'algorithme AdaBoost trois fois en parallèle sur les trois classifications binaires (convectif/non-précipitant ; convectif/stratiforme ; stratiforme/non-précipitant) effectuées par des SVM binaires composant cette variante OvO-SVM.

Dans la figure IV.4b, la combinaison de l'algorithme AdaBoost avec le multi-classifieur 'One versus All SVM' est illustrée. Trois classifications binaires (convective/tout ; stratiforme/tout ; non-précipitant/tout) réalisées en parallèle à l'aide de SVM binaires auxquelles l'algorithme AdaBoost a été appliqué pour chacune en sortie.

Dans le cas de la figure IV.4c, nous montrons la combinaison de l'algorithme AdaBoost avec le multi-classifieur Slant Binary Tree-SVM. L'algorithme AdaBoost est appliqué à la première classification binaire (Non-précipitant/précipitant) effectuée par le classifieur binaire SVM. La classe précipitante est ensuite scindée en deux classes (Convective/Stratiforme) en utilisant un autre SVM binaire auquel l'algorithme AdaBoost est appliqué une seconde fois.

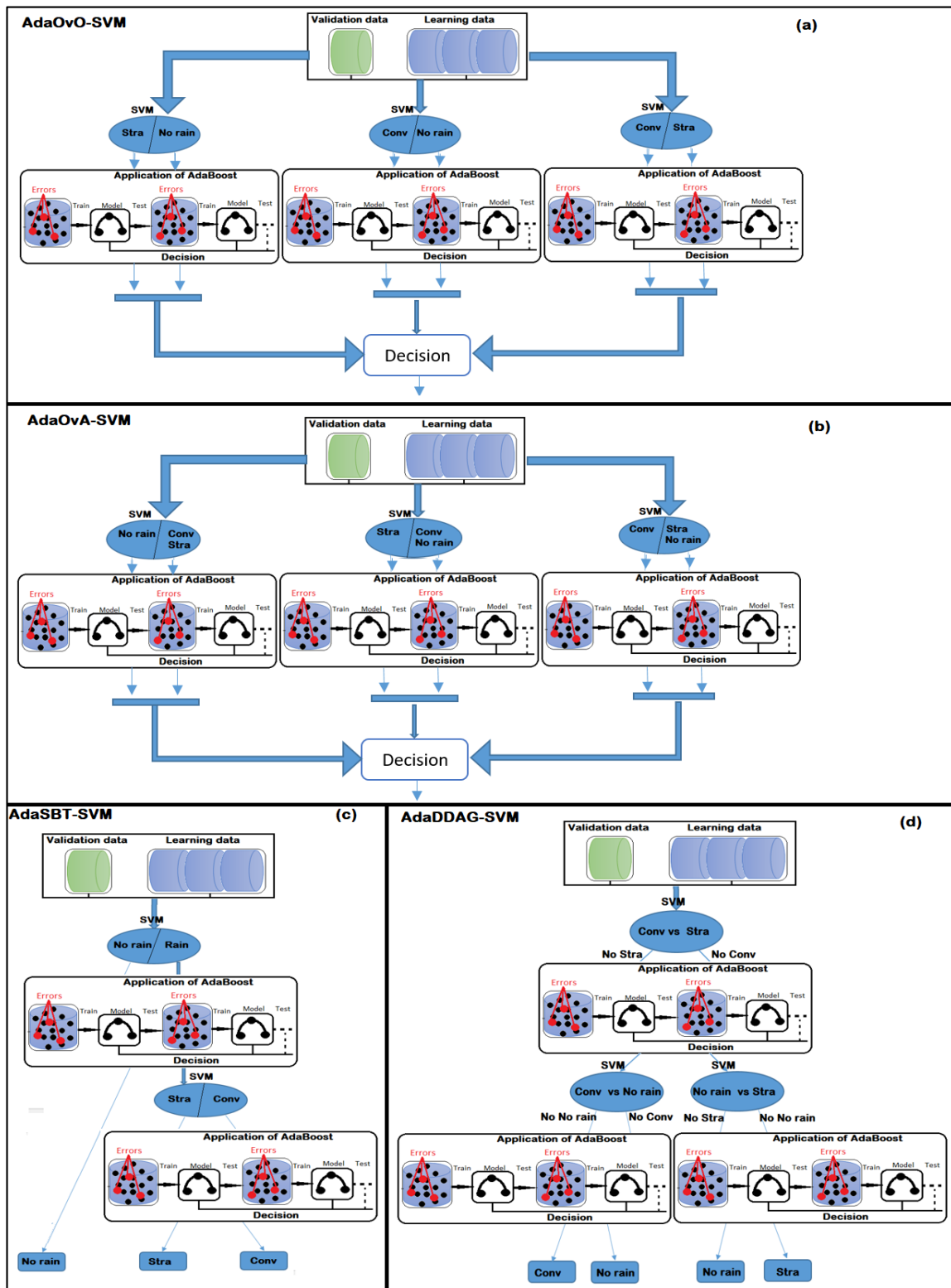


Figure IV.4 : Schéma d'application de l'algorithme AdaBoost sur les différentes variantes de SVM multiclassées, a) AdaBoost sur OvO-SVM, b) AdaBoost sur OvA-SVM, c) AdaBoost sur SBT-SVM, d) AdaBoost sur DDAG-SVM.

La figure IV.4d montre la combinaison de l'algorithme AdaBoost avec le multi-classifieur Decision Directed Acyclic Graph SVM (DDAG-SVM). L'algorithme AdaBoost est d'abord appliqué à la classification binaire (Non convective/Non stratiforme), obtenue par le premier SVM binaire composant cette variante. Ensuite, l'algorithme AdaBoost est appliqué deux fois en parallèle aux deux classifications binaires (non-stratiforme / non non-précipitant) et (non non-précipitant / non convectif) obtenues par les deux SVM binaires de la deuxième couche de ce modèle DDAG-SVM.

Pour mettre en œuvre le modèle AdaBoost-SVM développé pour toutes les variantes de cette étude, nous avons utilisé une base de données composée de trois saisons des pluies (section 2.2) qui ont été utilisées pour l'apprentissage des différents modèles SVM-multiclassés ainsi que l'apprentissage d'AdaBoost appliqué à ces modèles SVM-multiclasses, puis à l'application et tests des modèles résultants de l'application de AdaBoost aux différentes variantes SVM-multiclassées.

Le tableau IV.4 donne la taille des ensembles de données (échantillons) utilisés lors de l'apprentissage et lors de la validation des schémas développés, et le nombre de pixels par classe pour les deux périodes (apprentissage et validation).

	Période d'apprentissage						Période de Validation		
	Saison des pluies (2008-2009)			Saison de pluies (2009-2010)			Saison des pluies (2010-2011)		
	1799 IPS			2080 IPS			1287 IPS		
Classe	Convective	Stratiforme	Non-précipitant	Convective	Stratiforme	Non-précipitant	Convective	Stratiforme	Non-précipitant
Pixels observés par le radar	241044	874452	3541165	296542	986855	3805477	195745	745210	3021456
SVM	Apprentissage			Application			Classification		
AdaBoost	Non utilisé			apprentissage			Optimisation de la classification		

Tableau IV.4 : Périodes des bases de données d'apprentissage et de validation avec le nombre de scènes de précipitations instantanées (IPS) et de pixels par classe pour chaque période.

Une scène de précipitation instantanée (IPS pour Instant Precipitation Scene) est une scène qui contient au moins un pixel détecté comme précipitant. Les situations correspondant au ciel clair ne sont pas prises en compte.

Le classificateur SVM a donc été appris en utilisant 241044 pixels convectifs, 874452 pixels stratiformes et 3541165 non-précipitant (non-pluvieux). Le AdaBoost a été appris en utilisant 296542 pixels convectifs, 986855 pixels stratiformes et 3805477 pixels non-précipitants. Le modèle Ada-SVM est testé sur une base de données composée de 195745 convectifs, 745210 stratiformes et 3021456 non-précipitants.

IV.5. Tests et résultats des classifications

Les différents modèles construits dans cette étude seront testés dans cette partie. Au total, pour faire des comparaisons et vérifier le potentiel d'AdaBoost à optimiser les classifieurs SVM-multiclasse, quatre schémas de classification ont été développés. Dans le cas des classifieurs avec AdaBoost, nous fixons le nombre d'itérations à 400 pour éviter le sur-apprentissage comme illustré par le (tableau IV.5) et (figure IV.6).

Les modèles sont appliqués sur une scène de précipitation instantanée. Ensuite, nous avons appliqués les modèles développés sur toutes les scènes de précipitations de la période de validation.

IV.5.1. Application sur une scène instantanée :

La scène de précipitation instantanée sur laquelle nous avons appliqué les schémas de classification décrits est prise lors du passage d'un cyclone au-dessus de la région d'étude. Le cyclone est chargé de précipitations convectives et stratiformes. La scène est collectée par le satellite MSG le 29 novembre 2010 à 12h00 UTC pm.

Pour montrer l'apport de l'algorithme AdaBoost, nous avons classé la scène des précipitations en utilisant les quatre variantes de SVM multiclassées sans AdaBoost (OvA-SVM, OvO-SVM, DDAG-SVM, SBT-SVM) et avec AdaBoost (AdaOvA -SVM, AdaOvO-SVM, AdaDDAG-SVM, AdaSBT-SVM). Les résultats de la classification en trois classes

(convective, stratiforme et non-précipitant) sont illustrés à la figure IV.5 et la précision de la classification est indiquée dans le tableau IV.5.

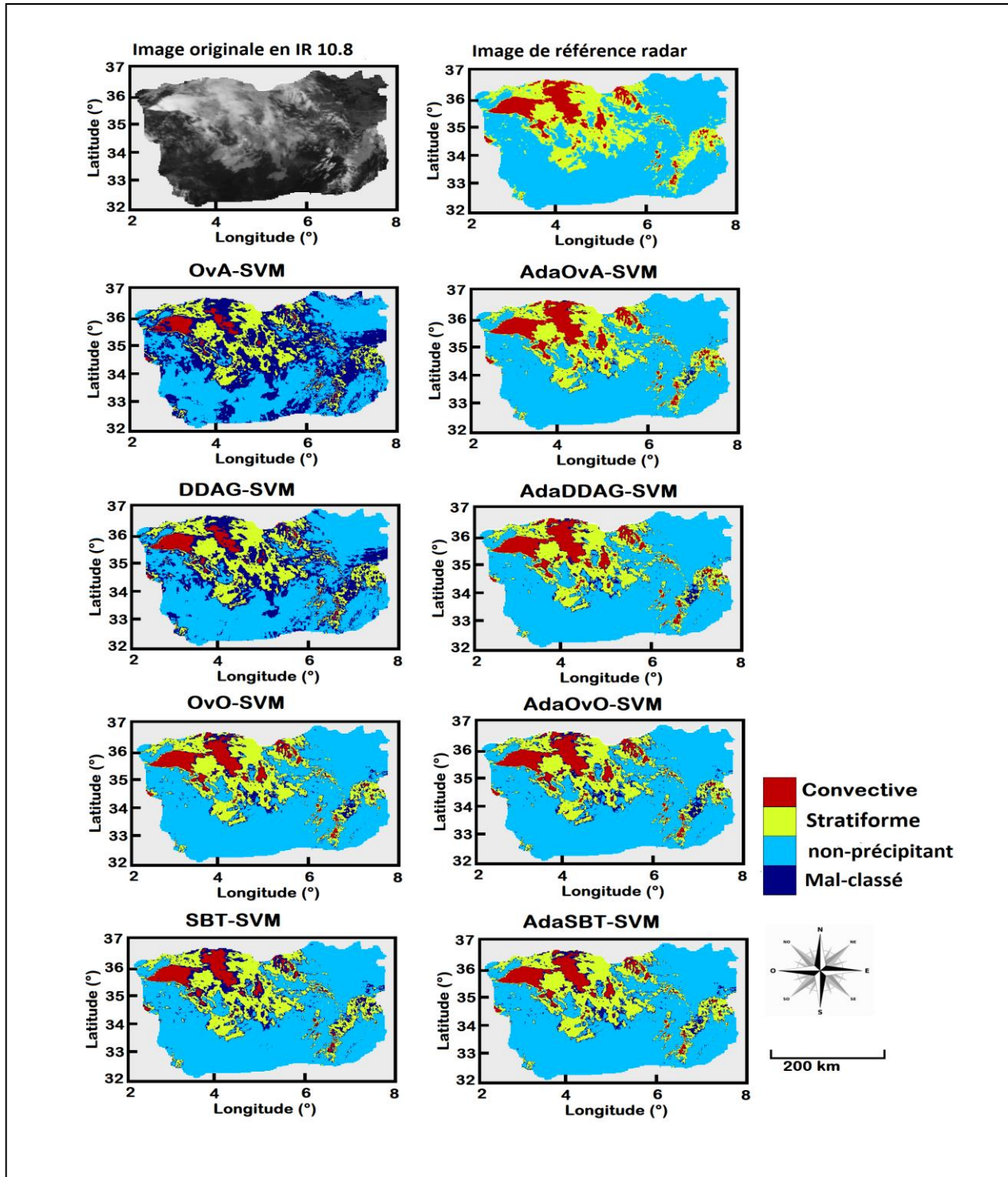


Figure IV.5 : Résultats de classification de la scène de précipitations instantanées (29 novembre 2010 à 12h00 pm), 1ère ligne : classée par radar (à droite), observée dans le canal infrarouge10.8 (à gauche). Entre la 2ème ligne et la 5ème ligne : résultats de classification sans AdaBoost (à gauche), résultats de classification avec AdaBoost (à droite).

Comme le montre la figure IV.5, les résultats de classification présentés indiquent que l'algorithme AdaBoost a bien fonctionné en optimisation pour certaines variantes, et n'a eu aucune amélioration ou à même eu une dégradation pour d'autres. En effet, un grand nombre de pixels mal classifiés obtenus pour le classifieur OvA-SVM (gauche, 2ème ligne) ont été correctement reclassifiés par AdaOvA-SVM (droite, 2ème ligne). Dans le cas de DDAG-SVM, l'algorithme AdaBoost a également montré son efficacité, où les pixels mal classifiés (gauche, 3ème ligne) sont reclassifiés correctement (droite, 3ème ligne). Quant aux modèles OvO-SVM et SBT-SVM, les résultats ne se sont pas améliorés en appliquant l'algorithme AdaBoost.

Pour montrer l'effet du nombre d'itérations AdaBoost sur la précision de classification, nous avons ainsi tracé les courbes représentant la précision de la classification en fonction du nombre d'itérations d'AdaBoost (0 à 412 itérations) pour toutes les variantes (Figure IV.6).

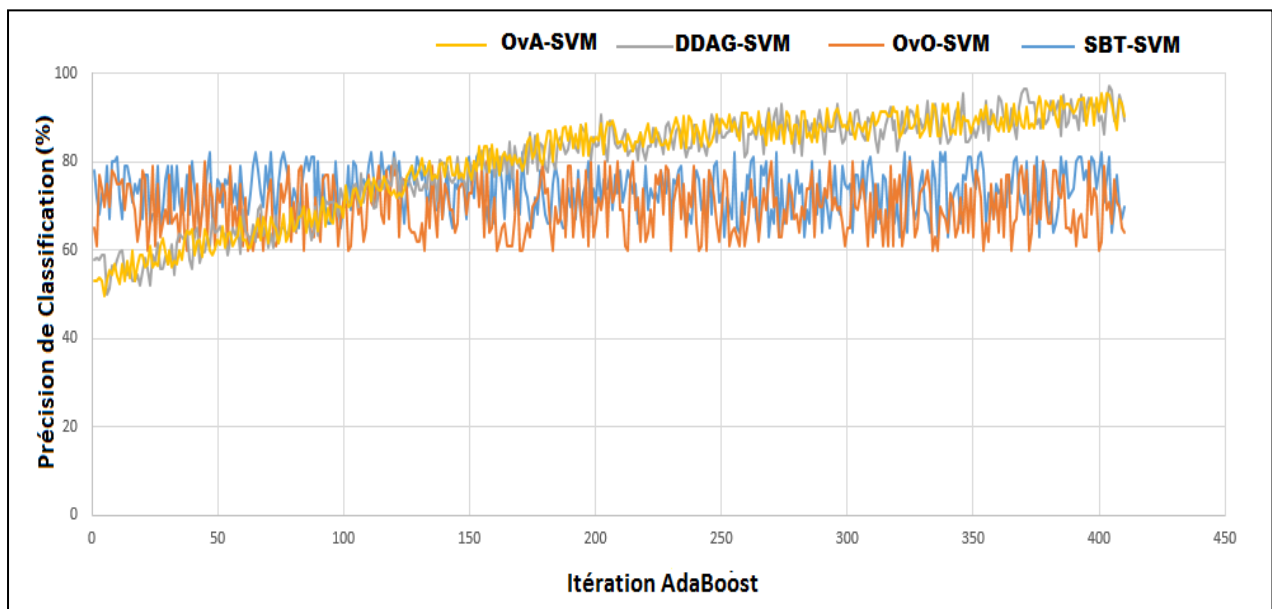


Figure IV.6 : Précision de la classification en fonction du nombre d'itérations d'AdaBoost.

Le tableau IV.5 montre les précisions de la classification en fonction des itérations 0, 100, 200, 300 et 400.

	Précision de classification (%)				
	0 itérations (SVMs sans AdaBoost)	100 itérations	200 itérations	300 itérations	400 itérations
AdaOvA-SVM	52.4	69.5	84.2	90.4	95.1
AdaOvO-SVM	70.2	75.1	74.2	68.4	69.2
AdaDDAG-SVM	58.2	65.3	82.4	88.6	90.1
AdaSBT-SVM	78.2	74.3	72.2	73.9	78.1

Tableau IV.5. Précision de classification après 0, 100, 200, 300 et 400 itérations d’AdaBoost pour les différentes variantes.

L'application de l'algorithme AdaBoost a considérablement réduit le nombre de pixels mal classifiés en les reclassifiant correctement pour les variantes OvA-SVM et DDAG-SVM. La précision de la classification passe de 52,4% à 95,1% à l'itération 400, passant par 84,2% à l'itération 200 pour OvA-SVM (voir Tableau IV.5). La même tendance a été observée pour DDAG-SVM, la précision de la classification démarre à 58,2% pour atteindre 90,1 % à l'itération 400.

Dans le cas des classificateurs OvO-SVM et SBT-SVM, AdaBoost n'a pas réussi à optimiser les résultats de la classification, même dans certains cas, les pixels étaient bien classifiés par OvO-SVM et SBT-SVM ont été reclassifiés de manière incorrecte par AdaBoost. Quelques pixels ont été reclassifiés correctement notamment pour la classe convective, mais la tendance générale (globale) montre que l'algorithme AdaBoost ne fonctionne pas pour les variantes OvO-SVM et SBT-SVM.

La précision de la classification passe de 70,2 % à 69,2 à l'itération 400 (tableau IV.5) pour OvO-SVM et de 78,2 % à 78,1 % pour SBT-SVM. Plus nous continuons à appliquer l’algorithme AdaBoost, plus la précision de la classification se stabilise.

Malgré quelques petites fluctuations, notamment aux itérations 110, 190, 210 et 300 où nous avons observé une remontée assez rapide de la précision, la courbe globalement suit un schéma ascendant, où le taux de bonne classification est proportionnel au nombre d'itération.

Cependant, il est connu qu'AdaBoost peut sur-apprendre les échantillons d'apprentissage et entraîner donc de mauvaises performances de généralisation. Il est donc nécessaire d'arrêter les cycles d'apprentissage d'AdaBoost au moment opportun à un nombre adéquat d'itérations, ainsi nous avons fixé la limite à l'itération 400. Comme nous pouvons l'observer, le sur-apprentissage commence à apparaître à partir de la 400^{ème} itération où le taux de bonne classification commence à descendre.

IV.5.2. Application sur toutes les scènes :

Dans cette section, nous présentons les résultats de la classification en appliquant les différents modèles construits sur toutes les scènes de précipitations de la saison des pluies 2010/2011. Le but est de montrer de manière objective la contribution apportée par l'algorithme AdaBoost dans le cas d'OvA-SVM et DDAG-SVM, mais aussi le cas où AdaBoost n'a pas bien fonctionné lorsqu'il est implémenté sur les variantes OvO-SVM et SBT-SVM. Pour ce faire, nous avons donc calculés des paramètres d'évaluation en utilisant les coefficients statistiques de la table de contingence (tableau IV.6), obtenus par les comparaisons avec les données réelles de radar météorologiques. Ces paramètres sont la probabilité de détection (POD : équation 4.17), la probabilité de fausse détection (POFD : équation 4.18), le taux de fausses alarmes (FAR : équation 4.19), l'indice de réussite critique (CSI : équation 4.20) et Pourcentage de correctes (PC : équation 4.21) et indice de biais de fréquence (biais : équation 4.22).

	Observé par le radar			
		Class(i)	No class(i)	Total
Classifié par le modèle	Class(i)	a	b	a+b
	No class(i)	c	d	c+d
	Total	a+c	b+d	a+b+c+d=n

Tableau IV.6 : Vue d'ensemble des combinaisons dans un tableau de contingence.

$$POD = \frac{a}{a+c} \quad (4.17)$$

$$POFD = \frac{b}{b+d} \quad (4.18)$$

$$FAR = \frac{b}{a+b} \quad (4.19)$$

$$Bias = \frac{a+b}{a+c} \quad (4.20)$$

$$CSI = \frac{a}{a+b+c} \quad (4.21)$$

$$PC = \frac{a+d}{n} \quad (4.22)$$

Pour les variantes de SVM-multiclasses avec et sans AdaBoost, nous obtenons les résultats statistiques présentés dans le tableau IV.7.

	La Classe Convective						La Classe Stratiforme					
	POD (%)	POFD (%)	FAR (%)	BIAS	CSI (%)	PC (%)	POD (%)	POFD (%)	FAR (%)	BIAS	CSI (%)	PC (%)
OvA-SVM	52.2	33.2	39.3	0.6	51.2	53.1	50.4	33.5	40.2	1.6	50.2	51.7
AdaOvA-SVM	95.2	12.4	14.7	0.9	88.1	96.5	94.4	12.9	15.5	1.2	87.8	96.1
OvO-SVM	88.2	18.5	22.3	0.6	86.2	89.2	84.5	20.0	23.9	1.6	83.8	85.0
AdaOvO-SVM	88.3	17.0	20.5	0.7	86.3	89.1	85.3	19.4	23.0	1.3	84.9	86.1
SBT-SVM	93.2	15.7	16.7	0.8	84.4	92.2	89.7	15.2	17.4	1.3	86.2	91.5
AdaSBT-SVM	91.4	16.2	17.3	0.8	83.4	92.1	90.4	16.8	19.3	1.4	85.3	90.9
DDAG-SVM	57.6	30.6	36.4	0.6	53.9	58.2	56.9	30.8	39.2	1.5	53.6	57.9
AdaDDAG-SVM	91.6	15.8	17.9	0.8	86.3	92.3	90.3	16.0	18.2	1.4	86.2	92.0
Optimal	100	0	0	1	100	100	100	0	0	1	100	100

Tableau IV.7. Paramètres d'évaluation de la classification des classes convectives et stratiformes pour différentes variantes de SVM multiclasses avec et sans AdaBoost.

Selon le tableau IV.7, la même tendance que les résultats de classification de la scène sélectionnée est observée pour la classification de l'ensemble des scènes. En effet, les performances des variantes OvA-SVM et DDAG-SVM ont été optimisées par l'implémentation de l'algorithme AdaBoost. Toutes les valeurs des paramètres d'évaluation indiquent une nette amélioration des classifications tant pour la classe convective que pour la

classe stratiforme. Pour les variantes SBT-SVM et OvO-SVM, qui ont fondamentalement des performances relativement élevées, la situation n'a pas changé avec l'application de AdaBoost, voire s'est légèrement dégradée. La variante qui a beaucoup bénéficié de l'incorporation de l'algorithme AdaBoost est OvA-SVM où ses performances se sont remarquablement améliorées.

En effet, pour la classe convective, les POD, POFD et FAR passent respectivement de 52,2%, 33,2% et 39,3% à 95,2%, 12,4% et 14,7%. Le biais indique une sous-estimation de 0,6 avant d'être ramené à 0,9 avec l'application d'AdaBoost. Quant au CSI et au PC, les valeurs passent respectivement de 51,2 % et 53,1 % à 88,1 % et 96,5 %. Et comme pour la classe convective, les résultats d'évaluation vont dans le même sens pour la classe stratiforme. Sans AdaBoost, les POD, POFD et FAR affichent respectivement 50,4%, 31,5% et 40,2%, avec AdaBoost, ces valeurs passent à 94,4%, 12,9% et 15,5%. Il y a une surestimation de 1,4 sans l'AdaBoost, elle a été réduite à 1,2 avec l'AdaBoost. En termes de CSI et de PC, la mise en place d'AdaBoost a amélioré ces paramètres. Le CSI passe de 50,2% à 87,8% et le PC passe de 51,7% à 96,1% .

Tous ces résultats prouvent que l'algorithme AdaBoost a corrigé les fausses classifications en l'implémentant sur le classifieur OvA-SVM, et un peu moins pour DDAG-SVM. Par contre, l'AdaBoost a montré des limites lorsqu'il est utilisé dans les classifieurs OvO-SVM et SBT-SVM.

Aussi, ce qui est important de noter, est que ces résultats dépassaient même ceux obtenus en appliquant les OvO-SVM et SBT-SVM sans AdaBoost, alors qu'au début on notait un écart considérable. Ces résultats montrent encore une fois que l'algorithme AdaBoost fonctionne bien sur les classifieurs OvA-SVM et relativement bien sur DDAG-SVM. Ces deux variantes ont montré des résultats mitigés au début.

La raison peut s'expliquer par le fait que pour ce type de données utilisé dans notre étude, il existe un fort déséquilibre entre les différentes classes. Des expériences sur des ensembles de données déséquilibrés, ont montré que AdaBoost avec SVM comme apprenant faible fonctionnait bien mieux que SVM sans AdaBoost (Li et al, 2005). Le OvA-SVM qu'est sensible à ce déséquilibre n'a pas réussi à faire une bonne classification au départ. Cependant, cet inconvénient a été exploité en considérant le OvA-SVM comme un classifieur (apprenant) faible pour le AdaBoost, et un peu moins pour le cas de DDAG-SVM, du moins pour ce type de données.

Dans le cas de ‘OvO-SVM’ et de SBT-SVM, les classifications sont correctement effectuées dès le départ. Ces variantes de SVM multiclassées ne peuvent pas être boostées par l'algorithme AdaBoost car elles ne sont pas considérées comme des classifieurs faibles.

Les variantes SBT-SVM et OvO-SVM sont moins sensibles dans le cas où les classes sont déséquilibrées. Pour cette raison, les performances de ces variantes n'ont pas été améliorées.

Pour répondre à la problématique concernant l'implémentation de AdaBoost sur SVM en le considérant comme apprenant faible et la possibilité que AdaBoost optimise le SVM, la réponse dépend du moins, selon cette étude, de deux paramètres : de la variante du classifieur SVM utilisée et de l'environnement où le classifieur est appliqué c'est à dire type de données utilisées et leurs caractéristiques.

IV.5.3. Inter-comparaison

Pour voir l'apport et la contribution du modèle développé dans notre étude, qui est l'AdaOvA-SVM par rapport à certains travaux déjà publiés dans la littérature, nous avons implémentés les techniques suivantes :

- La méthode ‘ Enhanced Convective Stratiform Technique (ECST)’: qui est similaire au CST(Convective Stratiform Technique) mais inclut en plus la température du canal de vapeur d'eau pour une discrimination plus fiable des nuages de convection profonde/cirrus (Tjemkes et al, 1997 ; Reudenbach, 2003 ; Reudenbach et al., 2001). La CST a été optimisé en considérant en outre le BTD (différence de température de brillance) entre un canal de vapeur d'eau et un canal infrarouge pour une identification plus fiable des nuages convectifs profonds et des zones de précipitation convective-stratiforme associées.
- La technique combinant trois classifieurs, à savoir ‘Support vector machine, Artificial neural network and Random forest Technique (SART)’ est conçue pour améliorer la classification des précipitations convective et stratiforme (Lazri et Ameer, 2018). Le réseau de neurones utilisé dans cette combinaison est de type MLP(Multi Layer Perceptron) avec une architecture optimisée composée de trois couches : entrée, cachée et sortie. Deux SVM binaires (Binary Tree Strategy) successifs et un modèle Random Forest (forêts aléatoire) avec 400 arbres (ntree) et 6 variables prédictives (mtry).

- Convective/Stratiform Rain Area Delineation Technique (CS-RADT) (Lazri et al, 2013) : est la technique de classification des précipitations par satellite, basée sur divers paramètres spectraux de SEVIRI, qui expriment les propriétés microphysiques et optiques des nuages. Elle utilise une technique de seuillage multispectrale pour distinguer entre les nuages stratiformes et convectifs. Les différents seuils ont été choisis par optimisation à partir des données de référence de Radar.
- La Random Forest Technique (RFT) (Ouallouche et al, 2018) est utilisée pour classifier les précipitations en trois classes (non précipitant, convective et stratiforme) en utilisant la théorie de la forêt aléatoire (Random Forest RF). La RFT est optimisée avec 400 arbres (ntree) et 6 caractéristiques ou variables prédictives (mtry), ceci en minimisant le taux d'erreur « out-of-bag » (OOB).

Les résultats statistiques globaux de l'application de ces différentes techniques sont présentés dans le tableau IV.8.

	POD	POFD	FAR	Bias	CSI	PC
AdaOvA-SVM	94.32	9.02	24.47	1.11	85.53	95.31
ECST	62.18	12.04	40.01	1.14	53.66	87.41
CS-RADT	68.22	11.47	37.56	1.12	59.32	88.89
RFT	90.32	9.22	26.87	0.88	76.89	93.40
SART	94.87	8.18	13.98	1.08	86.58	96.08
Optimal	100	0	0	1	100	100

Tableau IV.8. Résultats statistiques d'évaluation pour les différentes techniques.

La méthode de 'AdaOvA-SVM' montre des performances très intéressantes. Elle a dépassé de loin les performances des méthodes ECST et CS-RADT. Les résultats montrent également que la méthode AdaBoost-SVM est plus performante que la méthode RFT, elle a égalé la méthode SART où trois classifieurs basés sur l'apprentissage automatique sont combinés. Ceci démontre clairement que l'algorithme AdaBoost peut transformer un classifieur faible ou un classifieur fort opérant dans un environnement défavorable (exemple données de classes déséquilibrées), en un classifieur robuste et peut même avoisiner les performances des meilleurs classifieurs.

Pour donner une impression visuelle sur la contribution (apport) du modèle élaboré, nous présentons à la (figure IV.7) les résultats de classification en utilisant les différentes méthodes comparées y compris notre méthode AdaOvA-SVM, appliquées sur une scène de précipitations d'un cyclone prise le 29 Novembre, 2010 à 12:00 UTC pm par le satellite MSG.

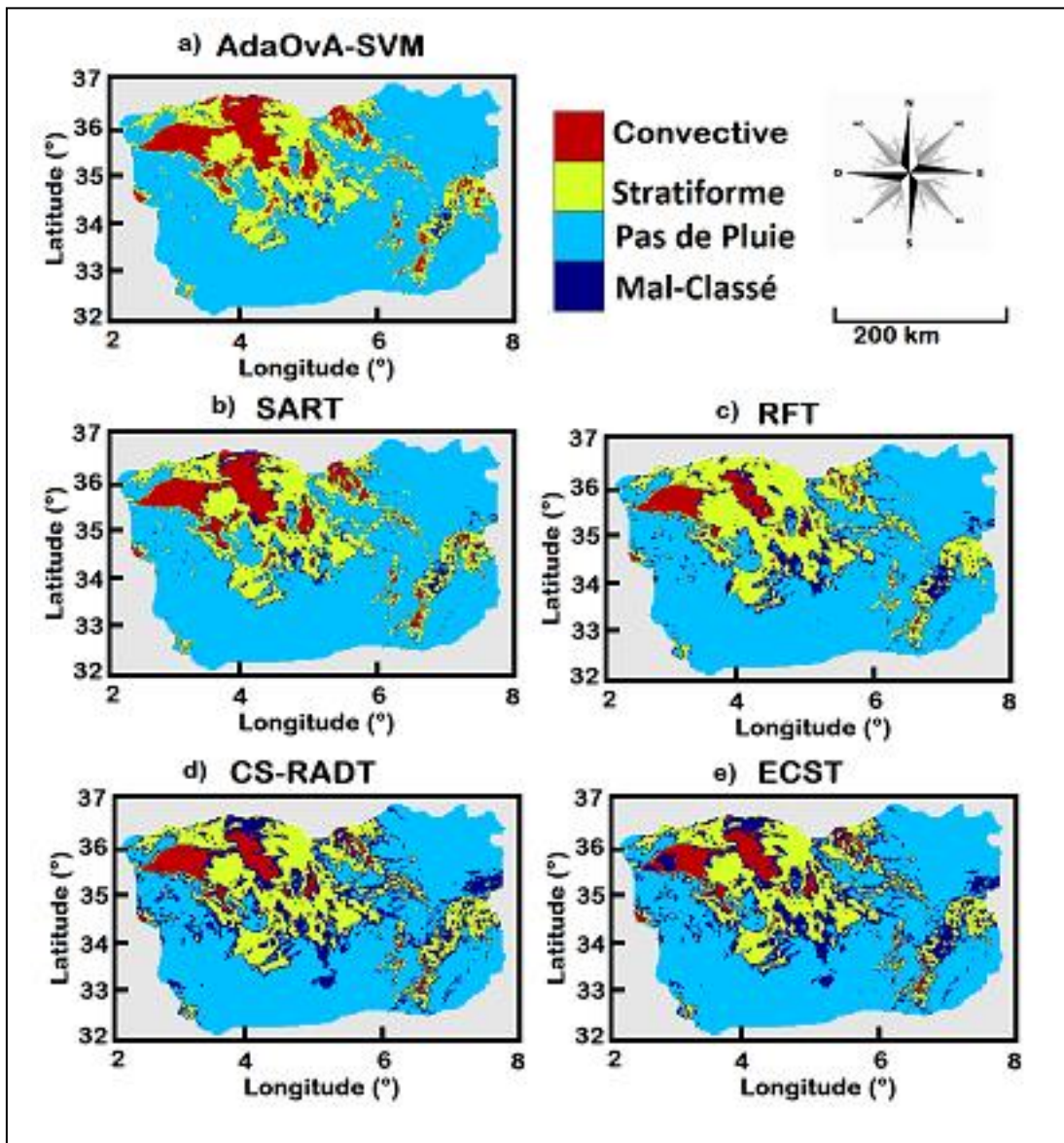


Figure IV.7. Résultats de la classification de la scène instantanée des précipitations (29 novembre 2010 à 12h00 pm), a) classée par AdaOvA-SVM, b) classée par SART, c) classée par RFT, d) classée par CS-RADT, e) classée par ECST.

Visuellement, les résultats de classification montrent que les pixels mal-classifiés sont moins présents dans le cas de AdaOvA-SVM par rapport aux méthodes ECST, CS-RADT et RFT. Les classifications obtenues par l'AdaOvA-SVM sont proches de celles obtenues par la méthode SART. Les pixels mal-classifié sont principalement présents aux frontières entre la classe convective et la classe stratiforme. La classe non précipitant est facilement classifiable.

IV.6. Estimation des taux de précipitations

Afin de situer les modèles développés dans notre travail, en ce qui concerne leurs capacités à bien estimer les taux de précipitations, par rapport aux mesures réelles des pluviomètres ou radar, nous avons dans cette section présenté les résultats à l'aide de la méthode satellitaire élaborée (AdaOvA-SVM).

Pour chaque pixel classifié par notre méthode, nous avons appliqué l'expression donnée par (l'équation 4.23) pour estimer le taux de précipitation de ce pixel:

$$RR(mm) = \frac{Nc \times Rc + Ns \times Rs}{4} \quad (4.23)$$

Nc et Ns représentent respectivement, le nombre d'apparition de la classe convective et le nombre d'apparition de la classe stratiforme respectivement, dans la classification d'un pixel au cours de la période de mesure.

Les images sont collectées à une fréquence de 15 minutes, permettant quatre classifications en une heure pour chaque pixel.

Rc et Rs représentent les intensités de précipitations convectives et stratiformes d'un pixel respectivement calculées à partir de comparaisons avec les données radar pour la saison des pluies 2009/2010. Les expressions sont données par les équations 4.24 et 4.25 :

$$Rc(mm/h) = \frac{CCP(mm)}{K_c} \times 4 = 4.2 mm/h \quad (4.24)$$

$$Rs(mm/h) = \frac{CSP(mm)}{K_s} \times 4 = 1.8 mm/h \quad (4.25)$$

CCP et CSP sont les précipitations convectives cumulées et les précipitations stratiformes cumulées enregistrées par radar, représentant les taux cumulés résultants respectivement des pixels précipitants convectifs et stratiformes durant la période de l'étude. Kc et Ks sont le nombre de pixels convectifs et le nombre de pixels stratiformes correspondant respectivement au CCP et CSP, classifié par la méthode satellitaire durant cette même période.

Nous avons considéré trois échelles pour l'estimation des précipitations, l'échelle quotidienne, l'échelle mensuelle et l'échelle saisonnière.

Pour l'échelle journalière, au cours de la journée du 29 novembre 2010, nous avons enregistré les données satellitaires d'un cyclone subtropical méditerranéen où une grande partie est constituée de précipitations convectives qui se sont formées à partir de 4h00 UTC AM et se sont totalement dissipées vers 20h45 UTC PM.

Dans le cas de l'échelle mensuelle pour le mois de novembre 2010, nous avons collecté des données satellitaires des événements de précipitations qui se sont produits au cours de ce mois. Nous avons noté la présence de pluies convectives et stratiformes.

La troisième échelle est l'échelle saisonnière. La saison 2010-2011 a connue différents types de précipitations où des événements convectifs et/ou stratiformes sont enregistrés.

En utilisant les résultats de classification obtenus en appliquant l'AdaOvA-SVM, nous avons estimé les taux de précipitations pour tous les pixels classifiés, en utilisant (l'équation 4.23) pour les trois échelles. Les estimations satellitaires des taux de précipitations, obtenue en utilisant la méthode AdaOvA-SVM, par rapport aux mesures pluviométriques sont représentées par la (Figure IV.8).

Le coefficient de corrélation entre les deux mesures est déterminé pour chaque échelle d'estimation en utilisant (l'équation 4.26).

$$r = \frac{\sum[(E_i - \bar{E}) * (V_i - \bar{V})]}{\sqrt{\sum(E_i - \bar{E})^2 * \sum(V_i - \bar{V})^2}} \quad (4.26)$$

Ou :

E_i : est la valeur du taux de précipitation estimée par la méthode satellitaire pour un pixel (i).

\bar{E} : La moyenne des valeurs estimées.

V_i : La vraie valeur mesurée du taux de précipitation pour le pixel (i).

\bar{V} : La valeur moyenne des taux de précipitations mesurées.

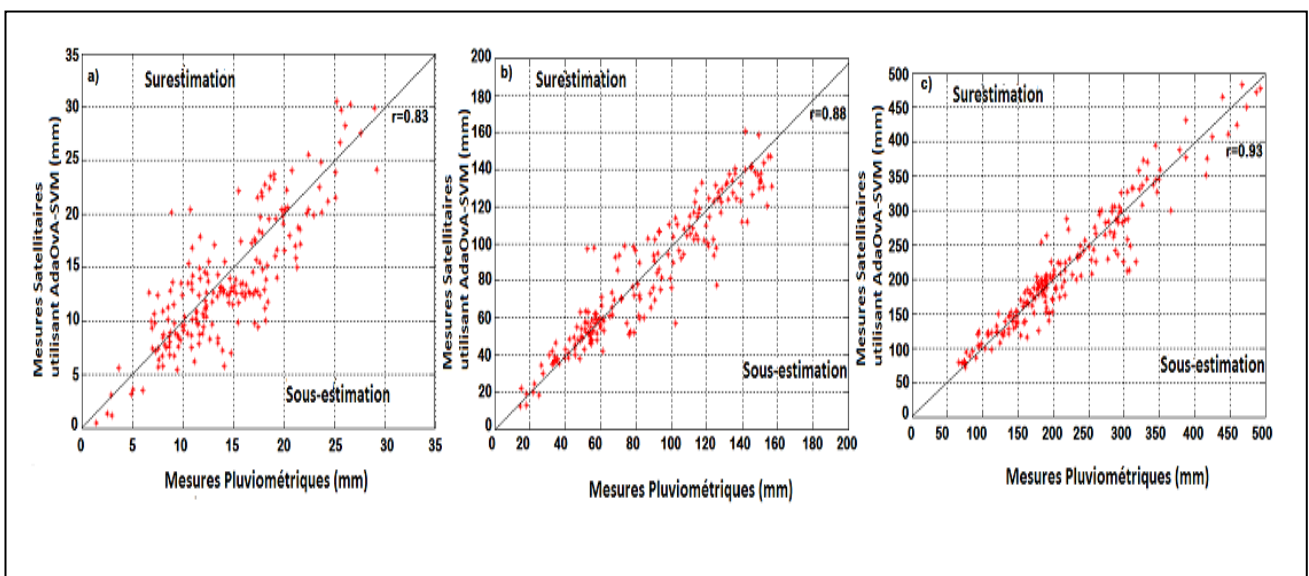


Figure IV.8 : Résultats de l'estimation des taux de précipitations utilisant AdaOvA-SVM par rapport au pluviomètre, a) échelle journalière, b) échelle mensuelle, c) échelle saisonnière.

La corrélation entre les estimations de la méthode satellitaire AdaOvA-SVM et les mesures pluviométriques (pluviomètre), est plus importante dans le cas de l'échelle saisonnière. La fluctuation du nuage de points dans ce schéma de corrélation est très annoncée dans le cas de l'échelle journalière. Cette situation s'est atténuée dans le cas de l'échelle mensuelle. Les coefficients de corrélation (0,83, 0,88, 0,93 pour l'échelle journalière, l'échelle mensuelle et l'échelle saisonnière, respectivement) montrent que les estimations à plus grande échelle s'équilibrent et la corrélation entre les deux types de mesure augmente, en raison de la succession et de la coexistence de plusieurs types de précipitations contrairement aux échelles plus courtes.

Pour évaluer la méthode développée AdaOvA-SVM en termes de performance dans les estimations des taux de précipitations, nous avons également estimé les précipitations en utilisant les méthodes décrites précédemment, à savoir ECST, CS-RADT, RFT et SART pour comparer leurs résultats à ceux de notre méthode.

Les métriques (paramètres d'évaluation statistiques) communes : erreur de biais moyenne (MBE ou biais), erreur absolue moyenne (MAE), racine de l'erreur quadratique moyenne (RMSE) utilisant les équations (4.27), (4.28) et (4.29) respectivement sont utilisées pour l'évaluation et la comparaison des résultats de ces méthodes.

$$MBE = \frac{1}{N} \sum_{i=1}^N (E_i - M_i) \quad (4.27)$$

$$MAE = \frac{1}{N} \sum_{i=1}^N |E_i - M_i| \quad (4.28)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (E_i - M_i)^2} \quad (4.29)$$

Où E_i et M_i sont respectivement la i ème estimation par méthode satellitaire et i ème mesure pluviométrique des taux de précipitations.

Les résultats des estimations des taux de précipitations à l'échelle saisonnière des différentes méthodes satellitaires, comparés aux mesures pluviométriques de précipitations sont calculés et illustrés dans la Figure IV.9 et leurs paramètres d'évaluation sont donnés dans le (Tableau IV.9).

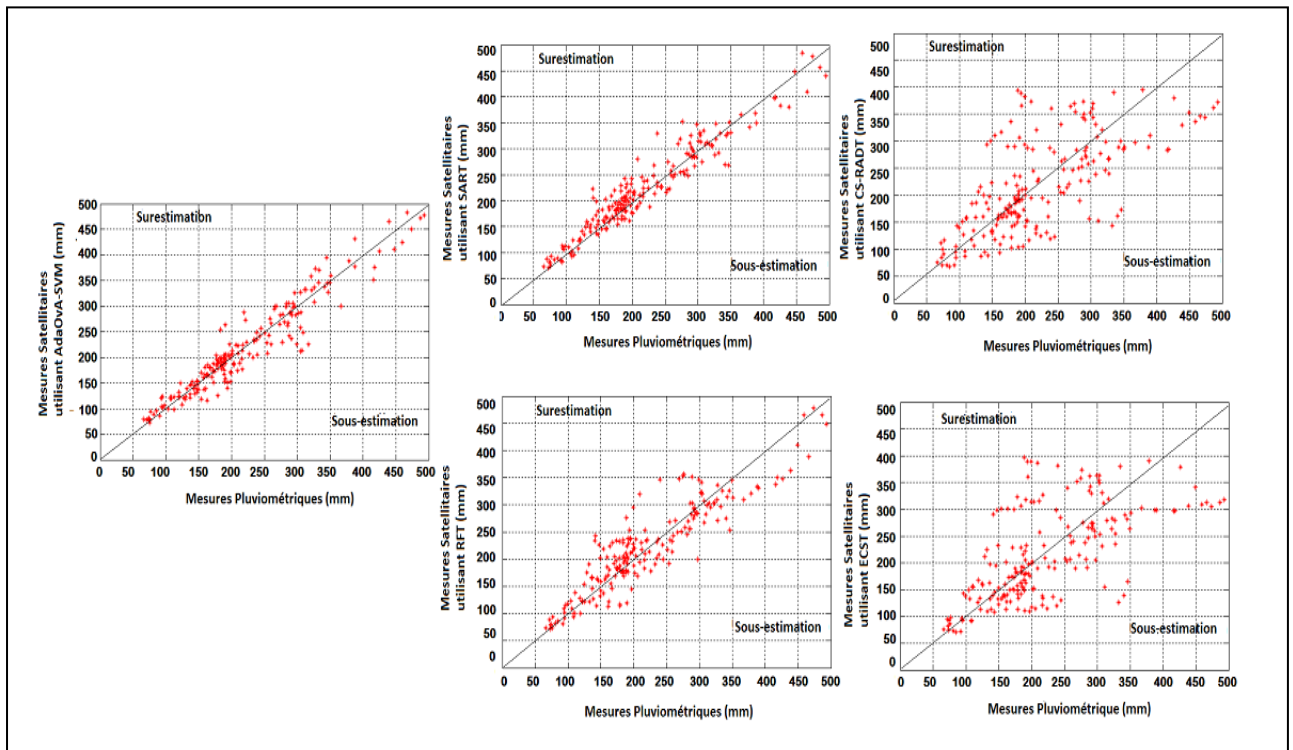


Figure IV.9 : Estimation des précipitations en utilisant les différentes variantes (AdaOvA-SVM, SART, RFT, CS-RADT et ECST) comparée aux mesures pluviométriques.

Comme pour la classification, les résultats des estimations de précipitations donnent une supériorité pour la méthode AdaOvA-SVM par rapport à l'ECST, CS-RADT, RFT.

L'AdaOvA-SVM, en termes de performances, est très proche de la méthode SART. Le coefficient de corrélation pour AdaOvA-SVM avec 0,93 indique une bonne corrélation avec les mesures de précipitations (figure9), la même corrélation est obtenue pour la méthode SART.

La technique développée est meilleure que les techniques ECST, CS-RADT, RFT en termes de corrélation avec les mesures réelles.

	MBE (mm)	MAE (mm)	RMSE (mm)	r
AdaOvA-SVM	6	25	45	0.93
RFT	-18	31	49	0.88
CS-RADT	26	37	57	0.84
ECST	33	40	64	0.76
SART	5	24	46	0.93

Tableau IV.9 : Paramètres d'évaluation des différentes méthodes dans le cas des estimations de précipitations.

En termes de MAE et RMSE, les paramètres d'évaluation des erreurs d'estimations (Tableau IV.9) montrent des performances intéressantes pour la méthode AdaOvA-SVM par rapport aux autres méthodes. Pour une quantité moyenne de précipitations de 238 mm, l'AdaOvA-SVM indique des valeurs de 25 mm et 45 mm pour ces deux paramètres

d'évaluation d'erreurs, contre 31 mm et 49 mm pour RFT, 37 mm et 57 mm pour CS-RADT, 40 mm et 64 mm pour ECST et, 24 mm et 46 mm pour SART confirmant ainsi la supériorité de AdaOvA-SVM.

Le Ada-OvA-SVM montre une légère surestimation des taux de précipitations, (MBE) de 6mm contre 26mm pour CS-RADT, 33mm pour ECST et 5mm pour SART, quant à la technique RFT, nous avons noté une sous-estimation de -18mm.

IV.7.Discussion

L'application dans cette étude de AdaBoost sur les différentes variantes de SVM-multiclasse, a permis de montrer l'existence de la possibilité d'utiliser AdaBoost avec SVM pour optimiser et améliorer ses performances, et de sélectionner la variante de SVM-multiclasse qui pourra être utilisée comme apprenant faible de l'algorithme AdaBoost.

En outre, l'exploitation de la technique de AdaOvA-SVM des informations sur les propriétés microphysiques et optiques des nuages telles que (les températures des sommets des nuages CTT, l'extension verticale CTH, et le contenu en eau liquide CWP, rendus disponible en exploitant les paramètres IR/VIS/WV de MSG-SEVIRI, a permis d'optimiser et améliorer la qualité et précision des classifications et l'estimation des précipitations.

En effet, le AdaOvA-SVM donne de meilleures précision de classification et estimations de précipitations par rapport à toutes les autres variantes avec AdaBoost, et même mieux que les résultats obtenus par les méthodes comparées.

Ces améliorations des résultats sont obtenues en appliquant AdaBoost sur la variante OvA-SVM de SVM-multiclasse, comparée aux autres variantes qui n'ont pas été significativement optimisées, dues au fait que la variante OvA-SVM a été sensible au déséquilibre des tailles des différentes classes. Ainsi, cette variante quand elle est utilisée seule, elle représente un classifieur faible. Ce qui a permis à l'algorithme AdaBoost de bien fonctionner sur le classifieur faible.

L'étude présentée dans ce chapitre répond à notre problématique fixée au début, à savoir la possibilité d'utiliser SVM comme classifieur faible, pour être optimisé par AdaBoost, à condition de trouver la variante adéquate de SVM-multiclasses.

Conclusion

L'objectif de ce travail est, d'une part, de réaliser une classification de qualité des intensités de précipitations dans le Nord de l'Algérie à partir d'images et données MSG, et d'autre part de montrer l'apport de l'algorithme AdaBoost sur le classifieur SVM multiclassés, et sa capacité à optimiser et améliorer la Précision de la classification SVM.

Pour ce faire quatre variantes principales de SVM multiclassés : OvA-SVM, OvO-SVM, SBT-SVM, DDAG-SVM ont été testées avec l'algorithme AdaBoost, pour trouver celle qui pourra être utilisée comme son apprenant faible. Des paramètres multi-spectraux de MSG-SEVIRI fournissant des informations implicites sur les propriétés optiques et microphysiques des nuages ont été sélectionnés et utilisés comme entrées du modèle.

Les différentes variantes de SVM multiclassée, ont été appliquées seule et avec introduction de AdaBoost, à une scène de précipitation instantanée d'un événement où des cellules convectives sont noyées (incrustées) dans un système stratiforme, dans le but de voir la quelle est plus adéquate à être utilisée avec AdaBoost. Afin de mettre en évidence cet apport d'AdaBoost, les différentes variantes de SVM multiclassée ont été appliquées dans un premier temps sans AdaBoost et dans un second temps, l'algorithme d'AdaBoost a été incorporé dans le traitement.

Les résultats de la classification montrent que le AdaBoost a réussi à optimiser la variante OvA-SVM et à un degré inférieur la DDAG-SVM. En effet, nous avons noté une nette amélioration de la qualité de la classification de cette variante après application de AdaBoost, les pixels mal classifiés ont été reclassifiés correctement en appliquant l'algorithme AdaBoost à l'itération 400, cette valeur optimale obtenue après application et tests de AdaBoost avec différents nombres d'itérations.

Cependant, les résultats étaient mitigés lorsque le AdaBoost est appliqué sur les variantes OvO-SVM et SBT-SVM. Ces deux variantes, ont montré des résultats relativement bons, dès le départ, avant l'application d'AdaBoost, ce qui fait que l'AdaBoost n'est pas recommandé pour un classifieur déjà robuste.

Dans le cas de classification sur toutes les scènes de précipitations, la même tendance est observée. Les paramètres d'évaluation ont montré que les meilleures classifications sont obtenues en utilisant la méthode AdaOvA-SVM. Ainsi, cette étude a

confirmée la puissance de l'algorithme AdaBoost lorsqu'il est combiné avec un classifieur faible, ou même sur un classifieur fort opérant dans un environnement défavorable, comme les classes avec des données déséquilibrées.

Nous avons aussi réalisé une inter-comparaison entre les résultats de classification obtenus par la méthode AdaOvA-SVM avec ceux des méthodes ECST, CS-RADT, RFT et SART. La méthode AdaOvA-SVM a bien montré des performances très intéressantes. Elle a dépassé les trois méthodes ECST, CS-RADT et RFT et ses résultats sont proches de la méthode SART où trois classifieurs d'apprentissage automatique ont été combinés.

Afin de situer le modèle développé, par rapport aux mesures réelles des pluviomètres et des taux de précipitations, nous avons réalisé des estimations de précipitations aux échelles journalière, mensuelle et saisonnière. Une meilleure corrélation des estimations réalisées par AdaOvA-SVM avec les mesures pluviométriques est obtenue dans le cas des estimations saisonnières. Toujours en termes d'estimation des taux des précipitations, les résultats montrent une performance remarquable pour AdaOvA-SVM par rapport à ECST, CS-RADT et RFT affichant la plus grande corrélation avec les mesures des pluviomètres. Cette performance est quasiment identique à celle du SART.

L'ensemble de ces résultats montre que le SVM peut être utilisé comme classifieur (apprenant) faible pour le AdaBoost, à condition de le tester sur le type de données utilisées et de trouver la variante adaptée de ce SVM.

En revanche, si l'implémentation de SVM donne des résultats avec un certain niveau de performance considéré correctes, dans ce cas, il n'est pas recommandé de l'utiliser comme classifieur faible pour AdaBoost.

De plus, dans la perspective de ce travail, il est intéressant de tester d'autres variantes de SVM avec AdaBoost, telles que Newton Support Vector Machine (NSVM), Proximal Support Vector Machine (PSVM) et Least-Squares Support Vector Machine (LS-SVM) pour identifier une autre variante candidate pouvant être utilisée avec AdaBoost comme son apprenant faible. La base de données, et la nature et distribution des données jouent également un rôle dans la sélection des variantes SVM candidates.

Annexe A : Types de précipitations et propriétés microphysiques et optiques des nuages

A.1. Précipitations dans la région méditerranéenne :

Comme nous avons vu les deux type principaux des précipitations à savoir (convective et stratiforme), nous nous intéressons aux précipitations dans la région méditerranéenne, qui connaît ses propres caractéristiques climatiques particulières influencée par les situations géographiques et topographiques de la région, et aussi par la mer méditerranéenne. Ce qui lui donne un climat à la fois influencé par les courants et vents froids des moyennes latitudes du nord et le climat subtropical du sud (Trigo et al 2006), tel que les cyclones du front polaire apportent froid et humidité alors que les pressions tropicales et subsahariennes font remonter de l'air chaud et sec. Ce qui fait que les climats méditerranéens sont des climats instables et d'une pluviométrie extrêmement variable, se composant d'un mixe de précipitations convective de méso échelle et de précipitations de grande échelle ayant les caractéristiques stratiforme en zone de couverture étendu et intensité modérée .

Dans cette région, les épisodes de pluie intense (HPEs, pour Heavy Precipitating Events) peuvent être causés soit par un système perturbé persistant de grande échelle, associé à un front froid soit par un système convectif de méso-échelle (MCS, pour Mesoscale Convective System) quasi-stationnaire. Tel que le premier système est à l'origine des pluies à intensité modérées sur une longue durée, tandis que le second produit des pluies à fortes intensités sur une courte durée de temps de quelques heures à une journée. Les deux systèmes peuvent coexister et contribuer aux forts cumuls des précipitations dans ces régions.

A.1.1.Systèmes précipitants de grande échelle :

Dans ce type de systèmes précipitants, les précipitations sont caractérisées par des intensités modérées (généralement détectées avec des réflectivités radar < 38 dBZ) mais qui peuvent contribuer à de grands cumuls dues à leur caractère continu et persistant dans le temps (de durée supérieure à 24 h).

Ces systèmes sont liés aux systèmes frontaux des moyennes latitudes, dans les quels les masses d'air chaud sont toujours soulevées en altitude par la masse d'air froid (Figure A.1).

Annexe A

Les systèmes fortement précipitants de grande échelle se génèrent dans des environnements très dynamiques, où sont générés les ascendances qui produisent des formations nuageuses et précipitantes.

Les précipitations liées à ce système de grande échelle sont principalement stratiformes, avec l'existence éventuelle de quelques cellules convectives incrustées dans la masse. Ces précipitations se caractérisent par leur grand étendu spatiale et leur intensité modérée. Leur contribution aux forts cumuls malgré leurs intensités modérées est due à leur persistance sur la même zone pendant une assez longue durée. Le relief de la région méditerranéenne renforce et contribue au développement des précipitations de ces systèmes de grande échelle c'est l'effet d'orographie. Selon Houze (1993), la présence d'un relief sur la trajectoire d'une masse d'air entraîne son élévation, ce qui accentue les ascendances dynamiques et permet la saturation en contenue en eaux d'une couche atmosphérique plus étendue, ce qui constituera un plus grand potentiel précipitant.

La (figure A.1) montre une illustration schématique des fronts froid et chaud.

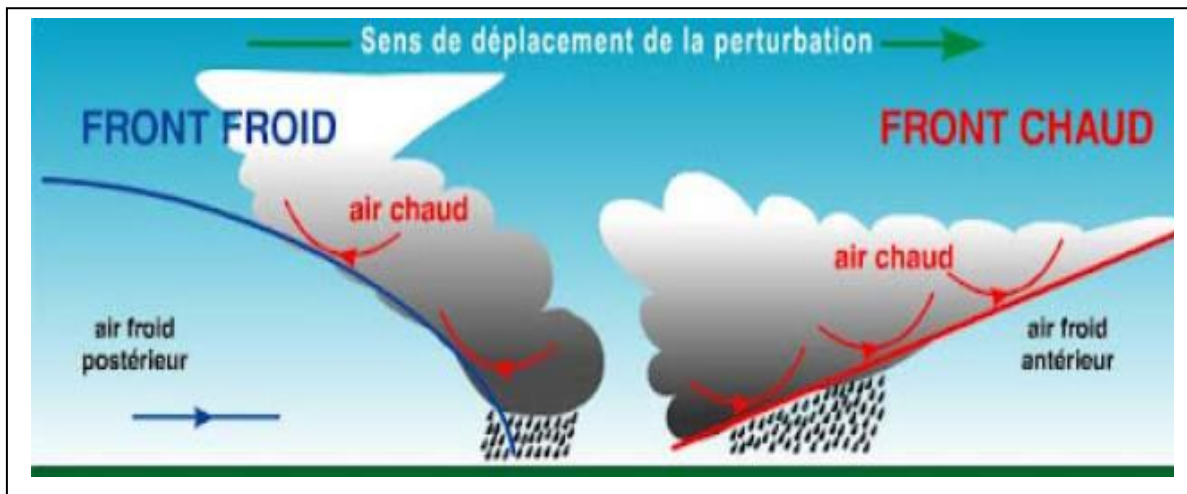


Figure A.1 : front froid et front chaud.

A.1.2. Systèmes précipitants de méso-échelle :

Les précipitations produites par les systèmes convectifs méso-échelle (MCS pour Mesoscale Convective System), sont présentes à différentes latitudes et avec des structures différentes, ils induisent les pluies intenses en Méditerranée. Les MCS peuvent être issues de

Annexe A

la combinaison de plusieurs systèmes convectifs de petite échelle tels que les orages monocellulaires, multicellulaires, ils sont donc composés de cellules convectives à différents stades de développement.

Généralement, un Système Convectif à Mésos-échelle est beaucoup plus grand qu'un orage individuel. Ce sont les MCSs qui sont les plus étendus des systèmes nuageux convectifs. Les vents associés aux MCSs sont des phénomènes locaux importants et particulièrement près des côtes du continent. Ils sont à l'origine de la grande majorité des épisodes de pluies intenses en région méditerranéenne (Riverain, 1997).

Les MCS se présentent sous plusieurs formes différentes. Cependant, elles peuvent toutes être caractérisées par un sommet très froid et verticalement très développé, reconnaissable grâce aux données des canaux infrarouge des satellites météorologiques. Chaque élément convectif a un cycle de vie; à la fin il s'affaiblit et devient une région de pluie stratiforme.

A.2. Propriétés radiatives des précipitations et des nuages

La connaissance des différentes natures d'interaction des nuages précipitants avec les émissions dans les différentes longueurs d'onde visible VIS, proche infrarouge NIR et infrarouge IR, due aux propriétés radiatives de ces nuages, qui dépendent de leurs propriétés optiques et microphysiques (épaisseur optique des nuages, rayon effectif des particules des nuages, phase thermodynamique des nuages, extension verticale des nuages), constitue des informations très importantes, pour mieux pouvoir détecter les différents nuages précipitants et d'estimer leurs contributions en précipitations, plusieurs études ont été menées dans l'optique de déterminer ces propriétés (Matrosov et al.1992 ; Mace et al.1998a ; Nakajima et King 1990) .

Selon les travaux de (Levizzani et al. 2001) ces propriétés des nuages selon les quels les méthodes vont estimer leurs potentiel précipitants, pouvaient être décrites comme suit :

- ✓ Dans l'infrarouge thermique, les propriétés radiatives sont sensibles à la distribution de la taille des hydrométéores. De telle sorte qu'une augmentation de la taille des particules augmente la transmissivité, diminue la réflectivité et augmente l'émissivité des nuages.
- ✓ L'émissivité des nuages de glace est inférieure à celle des nuages d'eau.
- ✓ Dans la bande (3.7-3.9 μ m), l'émissivité des nuages est plus faible que dans

Annexe A

l'infrarouge thermique (10.5 - 12.5 μm).

- ✓ Les nuages avec des petits hydrométéores diffusent et reflètent une grande partie du rayonnement dans la bande (3.7 -3.9 μm). Une augmentation de la taille des particules des nuages ou de la présence de grosses gouttes ou cristaux de glace au sommet du nuage réduit la réflectance dans la bande (3.7 -3.9 μm).
- ✓ La réflectance dans le NIR (1,50 - 1,78 μm) est inversement proportionnelle à la taille des rayons effectifs des particules de nuages.
- ✓ La réflectance dans la bande VIS est proportionnelle à l'épaisseur optique du nuage.

Parmi les paramètres des propriétés microphysiques et optiques des nuages, que les études ont essayé de déterminer, pour les utiliser dans la détections des nuages précipitants non convectifs, on trouve le rayon effectif des particules r_e et l'épaisseur optique des nuages τ .

Le rayon effectif (r_e) est défini par la relation suivante, selon (Levizzani et al. 2001) :

$$r_e = \frac{\int_0^{\infty} r^3 n(r) dr}{\int_0^{\infty} r^2 n(r) dr} \quad (\text{A.1})$$

Où r est le rayon des gouttelettes d'eau dans les nuages et $n(r)$ est la concentration de ces gouttelettes.

D'autres paramètres, qui sont la teneur en eau liquide des nuages LWC (Liquid Water Content) pour les nuages d'eau ou IWC (Ice Water Content) pour les nuages de glace, sont aussi utilisés pour caractériser les propriétés microphysiques des nuages et leur potentiel précipitant, ils sont définis selon [Levizzani et al.2001] comme suit:

$$LWC = \frac{4\pi \rho_w}{3} \int_0^{\infty} n(r) r^3 dr \quad (\text{A.2})$$

Où ρ_w est la densité de l'eau (en g.m^{-3}).

L'épaisseur optique des nuages et aussi l'un des facteurs utilisés dans ces méthodes, considérée pour détecter les nuages précipitants, elle est définie selon (Levizzani et al. 2001) par l'intégration du coefficient d'extinction des faisceaux par le nuage 'k' sur l'épaisseur géométrique de nuage:

Annexe A

$$\tau = \int_{z_1}^{z_2} k dz \quad (\text{A.3})$$

Où z_1 : est la hauteur de la base du nuage.

z_2 : est la hauteur du sommet du nuage.

et k : est le coefficient d'extinction ou atténuation des faisceaux par les nuages.

Le paramètre *LWP* (liquid water path) représente la quantité d'eau liquide intégrée sur la colonne atmosphérique (en $g \cdot m^{-2}$). Il s'exprime par la relation suivante (Levizzani et al. 2001) :

$$LWP = \int_{z_1}^{z_2} LWC(z) dz \quad (\text{A.4})$$

Le paramètre *CWP* (Cloud Water Path) constitue le *LWP* pour le nuage d'eau ou il représente le *IWP* (Ice Water Path) dans le cas des nuages de glace. Le *CWP* est proportionnelle à l'épaisseur optique des nuages (τ) et au rayon effectif des particules des nuages (r_e). C'est la grandeur utilisée plus souvent pour symboliser le contenu en eau des nuages et pour détecter les nuages précipitant non-convectifs. Il peut être exprimé en fonction de τ et de r_e par la relation suivante (Thies et al. 2008a; 2008b; Nauss et Kokhanovsky 2006):

$$CWP = \rho \cdot \frac{2}{3} r_e \cdot \tau \quad (\text{A.5})$$

Où : ρ ($\frac{g}{m^3}$) est la densité d'eau dans les nuages.

Un autre paramètre aussi très important dans les modèles utilisés pour la détection des nuages précipitants dans les zones non convectives (Thies et al. 2008a; 2008b; Lazri et al. 2013), est la phase thermodynamique des nuages (Cloud Phase CP).

Pour obtenir des informations sur cette phase thermodynamique des nuages, certains chercheurs ont utilisé les informations obtenues de la différence de températures de brillance des canaux SEVIRI qui est $\Delta T_{IR8.7-IR10.8}$ (Strabala et al., 1994; Thies et al., 2008b).

D'autres ont défini un seuil de température de brillance dans la bande IR10,8 pour faire la distinction entre la glace et l'eau des nuages (Rossow et Schiffer, 1999). Tandis que (Wolters et al., 2008) ont montré que l'utilisation simultanée de $T_{IR10.8}$ et $\Delta T_{IR8.7-IR10.8}$ permet une identification plus précise des phases thermodynamiques des nuages. Ils ont constaté que la formation de cristaux de glace commence lorsque le $T_{IR10.8}$ est supérieur à 238 K et le

Annexe A

$\Delta T_{\text{IR}8.7-\text{IR}10.8}$ est supérieur à 0,25 k. Il convient de noter qu'un nuage est susceptible de produire des précipitations lorsque des cristaux de glace sont présents au sommet.

Annexe B fonctionnement et équations du radar météorologique

B.1.Éléments constitutifs et principe de fonctionnement du radar météorologique :

Pour fournir les données utiles au domaine d'estimation des précipitations et la détection des contenus en eaux des nuages sondés, un système radar météorologique est constitué des éléments suivants : une antenne parabolique, un système d'émission/réception, un duplexeur et une unité de calcul et traitement (figure B.1) :

- ✓ **L'émetteur** : c'est la partie ou le composant responsable de la génération des impulsions électromagnétiques de courte durée, de forte puissance et haute fréquence qui sont émises par le radar.
- ✓ **L'antenne** : c'est le composant ou dispositif qui assure à la fois de la diffusion du signal dans l'espace, et de la réception du signal rétrodiffusé, car l'antenne radar est utilisée en émission/réception.
- ✓ **Le duplexeur** : utilisé dans ces cas du radar monostatique où un dispositif électronique commun et une même antenne réalisent la fonction d'émission et de réception du signal, ce duplexeur joue le rôle de commutateur en isolant le récepteur lors de la phase d'émission et vis-versa. En effet la puissance d'émission est souvent de l'ordre de plusieurs kilowatts, alors que le récepteur est conçu pour traiter quelques nano watts seulement.
- ✓ **Le récepteur** : c'est le composant qui s'occupe de la réception de l'onde et des signaux rétrodiffusés.
- ✓ **Une unité de calcul** : c'est la où les signaux subissent une phase de traitement avant l'affichage et la présentation des échos et produits radars obtenus.

La figure B.1 représente un schéma simplifié des composantes d'un système de fonctionnement du radar météorologique :

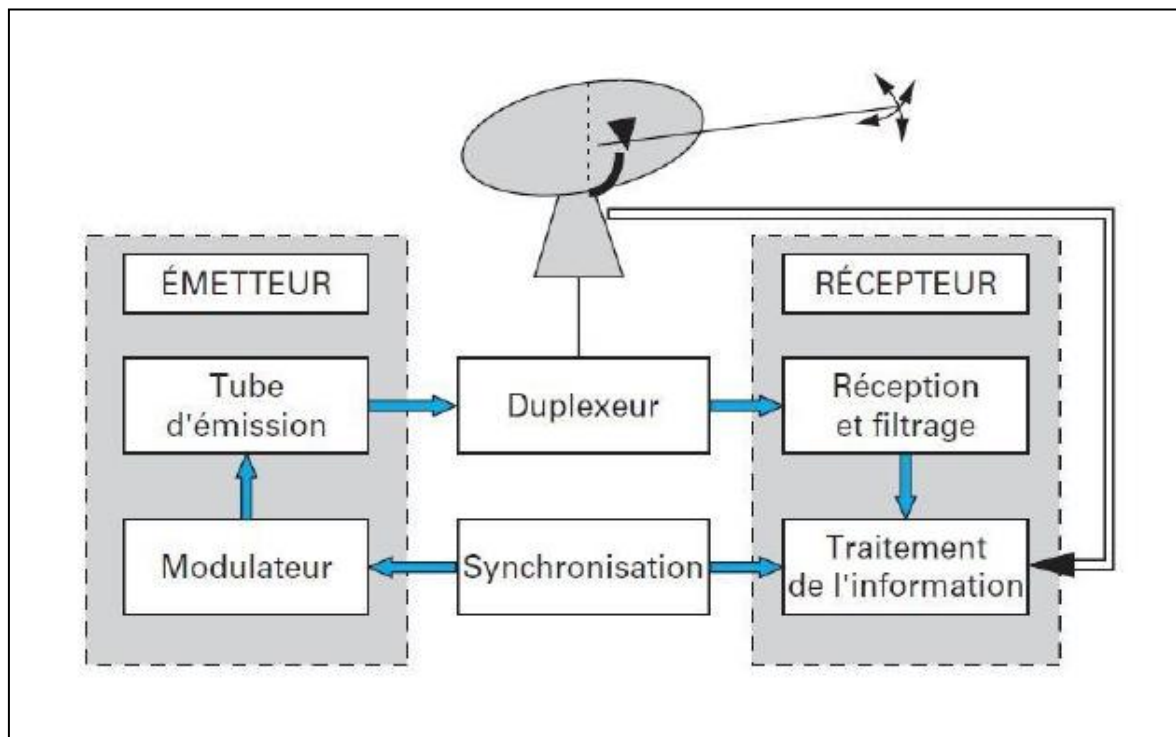


Figure B.1 : composants d'un système radar météorologique.

B.2.L'équation radar météorologique :

C'est une équation, qui exprime le rapport entre la puissance de l'onde émise par le radar et celle rétrodiffusée par la cible que sont les précipitations ou les nuages précipitants, exprimée en fonction des caractéristiques du radar, des conditions météorologiques et paramètres atmosphériques, ce qui montre les propriétés du volume diffusant d'une cible comme les nuages situés à une distance r du radar, cette distance qui est déduite à partir de l'écart de temps entre l'émission et la réception.

Il s'agit d'un bilan de puissance, entre l'émission et la réception du signal du radar météorologique après la rétrodiffusion par les nuages précipitants.

Considérons que P_T est la puissance totale émise par le radar, G_M le gain maximum de son antenne et σ la surface efficace de rétrodiffusion de la cible sondée, la distance ' r ' entre le radar et la cible est déduite du temps écoulé entre l'émission du signal et la réception du signal rétrodiffusé.

Annexe B

En négligeant l'effet de l'atténuation atmosphérique, cette équation radar qu'est le rapport entre la puissance émise et celle transmise est donnée par la relation suivante (Sauvageot,1992) :

$$\frac{P}{P_t} = \frac{\sigma G_M^2 \lambda^2}{(4\pi)^3 r^4} \quad (\text{B.1})$$

En tenant compte des atténuations L_a et L_r l'équation radar est de la forme suivante (Sauvageot, 1992) :

$$\frac{P}{P_T} = \frac{G_M^2 \lambda^2 \theta_0 \varphi_0 c \tau L_a^2 L_r}{1024 \pi^2 \ln 2} \cdot \frac{\eta}{r^2} \quad (\text{B.2})$$

Avec :

P : la puissance du signal rétrodiffusé reçue par le radar.

λ : Longueur d'onde d'émission du radar.

η : Réflectivité radar.

L_a : Facteur caractérisant l'atténuation de l'intensité du rayonnement dans l'atmosphère.

L_r : Affaiblissement lié à la bande passante de l'antenne du radar.

Et θ_0 et φ_0 les angles d'ouverture à 3 dB en site et en azimut respectivement.

La figure B.2 montre les différents paramètres de l'équation radar et de son fonctionnement :

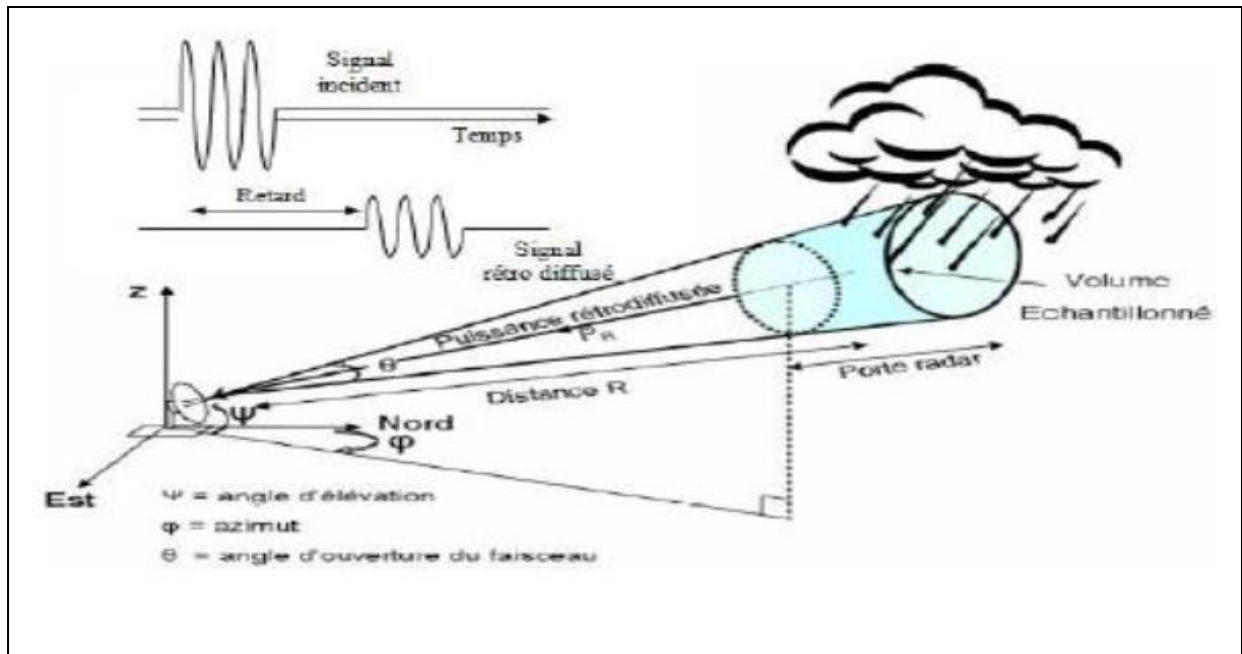


Figure B.2: les différents paramètres dans le fonctionnement du radar météorologique.

B.3. Résolution radiale d'un radar :

La résolution radiale d'un radar est sa capacité à distinguer entre deux cibles ponctuelles qui sont très proches l'une de l'autre, situées dans la même direction mais à des distances différentes. La résolution radiale dépend principalement de la largeur d'impulsion. Il est nécessaire que le système puisse différencier des cibles espacées d'un temps égal à une demi-largeur d'impulsion. En conséquence, la résolution radiale théorique d'un radar vaut (Sauvageot, 1992).

$$\Delta r = \frac{c\tau}{2} \quad (\text{B.3})$$

Avec :

C : vitesse de la lumière.

τ : Durée d'impulsion.

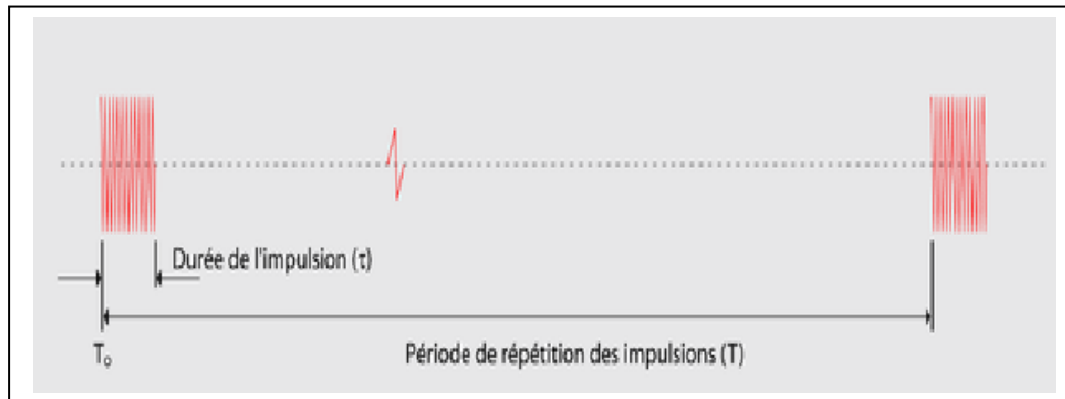


Figure B.3 : période et durée des impulsions radar.

B.4.Principaux types d'images Radar :

Les données ou images radar sont affichées dans un format qu'est déterminé par le mode de balayage de son antenne.

Plusieurs modes d'affichage sont utilisés, il s'agit des modes : PPI (Plan Position Indicator), RHI (Range Height Indicator) et HTI(Height Time Indicator).

✓ **Mode PPI :**

L'affichage radar PPI (Plan Position Indicator) est une représentation de la forme et de la position de la cible en azimut, dans ce mode une image radar est formée à partir d'un balayage scan à 360^0 en azimut à un angle d'élévation constant, l'antenne radar est généralement représentée au centre de l'écran de visualisation, de sorte que sa distance et sa hauteur par rapport au sol peuvent être représentées sous forme de cercles concentriques.

Lorsque l'antenne radar tourne, une trace radiale du PPI balaye l'unisson autour de son axe central, indiquant les positions en azimut dans ce plan des cibles.

✓ **Mode RHI :**

Les images radar obtenues à partir du mode RHI (Range Height Indicator) ,sont formées avec un balayage de l'antenne en azimute constant pour différentes angles d'élévation, ce mode d'affichage est intéressant pour visualiser la structure verticale détaillée d'un orage .

✓ **Mode HTI :**

Le mode d'affichage HTI(Height Time Indicator) est obtenu en effectuant des mesures sur une période de temps avec l'antenne orientée dans la direction verticale, Il s'agit d'une

Annexe B

mesure de l'intensité du signal émis et de son écho en fonction du temps, ce qui permet de déterminer la distance entre la cible et le radar.

B.5. Bandes de fréquences radar :

Les radars météorologiques utilisent différentes bandes de fréquences en fonction de leurs applications et de leurs cibles.

Les longueurs d'onde utilisées par ces radars sont récapitulées dans le (tableau B.1) ci-dessous, ainsi que leurs bandes de fréquences correspondantes et leurs catégories.

Bandes	Fréquences (Ghz)	Longueur d'onde (cm)	Les cibles détectées
S	2-4	7-15	Les précipitations/ les nuages en phase liquide / les nuages en phase de glace
C	4-8	4-7	
X	8-12	2.5-4	
Ku	12-18	1.7-2.5	
Ka	27-40	0.75-1.2	
W	90-100	0.3	Les nuages de glace

Tableau B.1 : Différentes bandes de fréquences et catégories des radars météorologiques.

Annexe C

C. Evaluation des performances des modèles d'apprentissage automatique supervisés :

La performance des modèles prédictifs d'apprentissage automatique dépend de plusieurs facteurs, comme les données utilisées pour la construction et l'apprentissage de ces modèles et le choix des paramètres d'optimisation et aussi la bonne adaptation du modèle choisit au problème à traiter.

Dans le cas de mauvais choix des métriques d'évaluation de la performance, qui devront être correctement adaptés par rapport à notre problème, ou de l'utilisation des méthodes et paramètres qui ne sont pas robustes, des conclusions peuvent être faites à tort que de bons modèles d'apprentissage automatique ont été construits, ce qui veut dire que la précision et la performance du modèle construit pourra être mal-estimée.

En apprentissage automatique, l'évaluation de la performance des modèles construits est une tâche à la fois cruciale et complexe.

Les métriques d'évaluation à utiliser pour cette fin dépendent de plusieurs facteurs, tels que la nature de la tâche de modélisation ou type d'apprentissage automatique à évaluer (classification, régression ou segmentation), l'ordre de grandeur des données ainsi que leur distribution.

Ces métriques d'évaluation, peuvent servir aussi à la comparaison des performances de plusieurs modèles d'apprentissage et sélectionner ceux qui donnent la meilleure performance.

Les parties suivantes abordent quelques unes des méthodes et des métriques qui sont utilisées pour tester et évaluer la performance prédictive des modèles d'apprentissage automatique supervisés (la classification, la régression).

C.1. Evaluation des performances et précisions des modèles de classification :

Pour les problèmes de classification et la vérification et l'évaluation de la précision des classifieurs, nous définissons la matrice de confusion, qui est utilisée dans le calcul des paramètres et scores d'évaluation de ce type d'apprentissage automatique.

Ensuite, ces mesures et paramètres d'évaluation qui sont (précision, rappel, spécificité et taux des faux négatifs) seront présentés.

C.1.1.La matrice de confusion :

C'est un tableau de taille (n x n), pour visualiser les résultats de classification des modèles prédictifs comparé aux classes réelles des exemples(observations) pour les problèmes de classification (Cornuéjols et Miclet 2010), où 'n' est le nombre de classes aux quels les données sont classifiées.

Dans cette matrice on croise les classes cibles réelles avec les classes prédites obtenues par les méthodes de classification, ce qui nous donne le nombre d'instances qui sont classées correctement et le nombre d'instances mal classées.

C'est une matrice M contenant un nombre de lignes et de colonnes identique au nombre de classes, et dont l'entrée ou l'élément M_{ck} représente le nombre d'exemples de la classe 'k' attribués à la classe 'c' par la méthode de classification.

Le nombre d'erreurs de classification permet d'évaluer la qualité d'un modèle prédictif.

		Classe réelle		
		0	1	
Classe prédites	0	Vrai négatif (TN)	Faux Négatif (FN)	Valeur prédictive positive (VPP)
	1	Faux Positif (FP)	Vrai Positif (TP)	Valeur Prédictive Négative (VPN)
		Rappel	Spécificité	Taux de succès= $\frac{TN + TP}{TN + FN + FP + TP}$

Tableau C.1: la matrice de confusion.

Dans les opérations de classification binaire (a deux classes) , par convention on choisit une classe comme la classe positive , comme exemple dans le cas de détection de nuages précipitants, ou on a les classes (précipitant/non-précipitant) , on considère la classe ' précipitant' comme positive et la classe 'non précipitant' comme négative .

Pour les différentes cases de la matrice de confusion :

- ✓ **TP (True Positive) en anglais ou (VP vrai positive)** : c'est le nombre des vrais positifs, le nombre d'instances appartenant à la classe positive correctement classifiées.
- ✓ **FP (False Positive)** : c'est le nombre des faux positifs, le nombre d'instances qui ne sont pas positives et qui sont prédites comme positives par la méthode de classification, les 'faux positifs' sont aussi appelés fausses alarmes.
- ✓ **FN (False Negatives)** : c'est le nombre des faux négatifs, le nombre d'instances non négatives (positives) classifiées comme négatives par la méthode.
- ✓ **TN (True Negative) ou Vrai Negative** : c'est le nombre des vrais négatifs, le nombre d'instances appartenant à la classe négative correctement classifiées.

C.1.2. Les scores de la matrice de confusion et leurs significations :

A partir de la matrice de confusion, on peut calculer plusieurs métriques utilisées pour l'évaluation des modèles de classification, et parmi ces métriques :

- ✓ **le taux de succès (Accuracy ou Hit ratio)** : C'est la proportion ou pourcentage des instances qui sont correctement classifiées.

$$A = \frac{TN + TP}{TN + FN + FP + TP}$$

- ✓ **le taux d'erreur (E)**: c'est l'erreur globale de la classification.

E = 1 - Taux de succès.

- ✓ **Rappel** : On appelle rappel, ou sensibilité (sensitivity en anglais), le taux de vrais positifs, c'est-à-dire la proportion d'exemples positifs correctement identifiés comme tels, par rapport à la totalité des exemples positifs.

$$\text{Rappel} = \frac{TP}{TP + FN}$$

Il est cependant très facile d'avoir un bon rappel en prédisant que *tous* les exemples sont positifs, donc ce paramètre a lui tout seul ne peut pas être pris en compte comme paramètre d'évaluation d'un modèle, Ainsi on lui adjoint souvent la *précision*.

- ✓ **Précision** : On appelle *précision*, ou *valeur positive prédictive* (*positive predictive value*, *PPV*) la proportion de prédictions correctes parmi toutes les prédictions positives :

$$\text{Précision} = \frac{TP}{TP+FP}$$

Pour résumer rappel et précision en un seul nombre, on calculera la *F-mesure* :

- ✓ **F-mesure** : On appelle *F-mesure* (*F-score* ou *F1-score* en anglais) la moyenne harmonique de la précision et du rappel :

$$F = 2 \cdot \frac{\text{Précision} \cdot \text{Rappel}}{\text{Précision} + \text{Rappel}} = \frac{2TP}{2TP + FP + FN}$$

- ✓ **Spécificité** : On appelle *spécificité* le taux de vrais négatifs, autrement dit la proportion d'exemples négatifs correctement identifiés comme tels.

$$\text{Spécificité} = \frac{TN}{FP + TN}$$

Notons que dans la littérature anglaise du domaine, on distingue précision (la précision décrite ci-dessus) et accuracy, qui est la proportion d'exemples correctement étiquetés, soit le complémentaire à 1 du taux d'erreur, aussi traduit par *précision* en français. On utilisera donc ces termes avec précaution.

Le taux de succès (Accuracy en anglais) est la métrique la plus utilisée dans l'évaluation de performance des modèles de classification. Toutefois, ce score n'est pas représentatif à lui seul de la mesure de l'efficacité et précision d'un modèle, il doit être combiné avec d'autres mesures et scores d'évaluation pour obtenir une évaluation fiable du modèle, surtout quand les classes des données utilisées ne sont pas uniformément distribuées (équilibrées).

C.2. Evaluation des performances des modèles de régression :

Dans le cas des problèmes de régression, où les sortie des prédictions sont des valeurs et mesures quantitative continue, le nombre d'erreurs ou d'estimations correctes ne sont pas des critères appropriés pour évaluer la performance.

En général un modèle qui donne globalement des prédictions plus proches des vraies valeurs pour la majorité des exemples est préféré à un modèle dont les prédictions sont presque exactes pour quelques points mais très loin des vraies valeurs pour beaucoup d'autres.

Ainsi, il est préférable dans ce genre d'évaluation de quantifier et juger la performance d'un modèle de régression en fonction de l'écart entre les valeurs prédites et les valeurs réelles, où on privilégie la moyenne des erreurs faite sur la totalité des données.

Et parmi ces paramètres qui mesure l'évaluation d'une régression étant données 'n' valeurs réelles : (y_1, y_2, \dots, y_n) et 'n' prédictions $f(x_1), f(x_2), \dots, f(x_n)$ on a :

C.2.1. La somme des carrés des résidus : ou **RSS** pour *Residual Sum of Squares*.

Pour calculer ce paramètre, il suffit de calculer pour chaque point x_i du jeu de test le carré de la distance entre sa vraie valeur y_i et la valeur prédite par le prédicteur f et en faire la somme.

$$RSS = \sum_{i=1}^n (f(x_i) - y_i)^2$$

C.2.2. L'erreur quadratique moyenne ou **Mean Squared Error (MSE)** :

On obtient l'erreur quadratique moyenne, en normalisant par le nombre de points total dans le jeu de test qui est 'n' la valeur RSS, car le problème de la RSS c'est qu'elle est d'autant plus grande qu'on a de données.

$$MSE = \frac{1}{n} \sum_{i=1}^n (f(x_i) - y_i)^2$$

C.2.3. Racine de l'erreur quadratique moyenne (**Root Mean Squared Error RMSE**):

Pour se ramener à l'unité de la mesure y ou au même ordre de grandeur que la mesure y on la racine de la MSE est calculée. On obtient ainsi la **RMSE**, ou *Root Mean Squared Error*.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (f(x_i) - y_i)^2}$$

C.2.4. Racine du logarithme de l'erreur quadratique moyenne (RMSLE) :

Dans le cas où les valeurs cibles y sont d'ordres de grandeur très variés, il est préférable parfois de passer au 'logarithme' avant de comparer $f(x_i)$ à y_i , pour éviter de donner à tort plus d'importance aux erreurs faites pour des valeurs d'ordre de grandeurs plus élevés, on obtient ainsi la 'Racine du logarithme de l'erreur quadratique moyenne (RMSLE)' :

$$RMSLE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\log(f(x_i) + 1) - \log(y_i + 1))^2}$$

C.2.5. Coefficient de détermination (Relative Squared Error RSE) :

L'interprétation de ces erreurs nécessite la connaissance de la distribution et la nature des valeurs cibles (étiquettes). Pour cela, il est possible de normaliser la somme des carrés des résidus, en la comparant à la somme des distances des valeurs cibles y_i à leurs moyenne.

$$RSE = \frac{\sum_{i=1}^n (y_i - f(x_i))^2}{\sum_{i=1}^n (y_i - \frac{1}{n} \sum_{i=1}^n y_i)^2}$$

C.2.6. Coefficient de corrélation(R) :

On a le complémentaire à 1 de la (RSE) le *coefficient de détermination* qui est noté R^2 , il s'agit du carré du coefficient de corrélation R.

Ce coefficient de corrélation R indique à quel point les valeurs prédites $f(x_i)$ sont corrélées aux valeurs réelles y_i .

$$R = \frac{\sum_{i=1}^n (y_i - \frac{1}{n} \sum_{i=1}^n y_i) (f(x_i) - \frac{1}{n} \sum_{i=1}^n f(x_i))}{\sqrt{\sum_{i=1}^n (y_i - \frac{1}{n} \sum_{i=1}^n y_i)^2} \sqrt{\sum_{i=1}^n (f(x_i) - \frac{1}{n} \sum_{i=1}^n f(x_i))^2}}$$

Bibliographie

Adler, R.F., Negri, A.J., 1988. A satellite infrared technique to estimate tropical convective and stratiform rainfall. *J. Appl. Meteorol.* 27, 30–51.

Arking, A., Childs, J.D., 1985. Retrieval of cloud cover parameters from multispectral satellite images. *J. Appl. Meteorol.* 24, 322–333.

Alpert P., Baldi M., Ilani R., et al. 2006: Relations between climate variability in the Mediterranean region and the tropics: ENSO, South Asian and African monsoons, hurricanes and Saharan dust, *Mediterranean Climate Variability, Elsevier B. V.*, 149–177.

Amorati, R., Alberoni, P. P., Levizzani, V., & Nanni, S., 2000. IR-based satellite and radar rainfall estimates of convective storms over northern Italy. *Meteorological Applications*, 7(1), 1-18.

Arkin P. A., 1979. The relationship between fractional coverage of high cloud and rainfall accumulations during GATE over the B-scale array. *Monthly Weather Review*, Vol 107, pp 1382-1387.

Arkin, P.A. and Meisner, B.N. 1987. The Relationship between Large Scale Convective Rainfall and Cold Cloud over the Western Hemisphere during 1982-84. *Monthly Weather Review*, 115, 51-74. [https://doi.org/10.1175/1520-0493\(1987\)115<0051:TRBLSC>2.0.CO;2](https://doi.org/10.1175/1520-0493(1987)115<0051:TRBLSC>2.0.CO;2)

Belghit, A., Lazri, M., Ouallouche, F., Labadi, K., & Ameer, S. 2023. Optimization of One versus All-SVM using AdaBoost algorithm for rainfall classification and estimation from multispectral MSG data. *Advances in Space Research*, 71(1), 946-963.

Belmahdi, F., Lazri, M., Ouallouche, F., Labadi, K., Absi, R., & Ameer, S. 2023. Application of Dempster-Shafer theory for optimization of precipitation classification and estimation results from remote sensing data using machine learning. *Remote Sensing Applications: Society and Environment*, 29, 100906.

Bensafi, N., Lazri, M., & Ameer, S. 2019. Novel WkNN-based technique to improve instantaneous rainfall estimation over the north of Algeria using the multispectral MSG SEVIRI imagery. *Journal of Atmospheric and Solar-Terrestrial Physics*, 183, 110-119.

Bibliographie

Breiman, L., Friedman, J., Olshen, R. and Stone, C. 1984. Classification and Regression Trees. Chapman and Hall, Wadsworth, New York.

Breiman, L. 1996. Bagging predictors. *Machine learning*, 24, 123-140.

Breiman, L. 2001. Random forests. *Machine learning*, 45(1), 5-32.

Breiman, L., 1998. Arcing classifiers (discussion paper). *Ann. Stat.* 26,801–824.

Breiman, L., 2000. Some infinity theory for predictor ensembles Technical Report 579. Statistics Dept UCB.

Brynjolfsson, E. and T. Mitchell. 2017. “What can machine learning do? Workforce implications”. *Science*. 358(6370): 1530–1534.

Carn M., D. Dagorne, B. Guillot et Lahuec J. P. 1989 : Estimation des pluies par satellite en temps réel en Afrique Sahélo-soudanienne. *Veille Climatique satellitaire*, N°28, pp 47-55.

Cheng, W.C., Jhan, D.M., 2013. Triaxial accelerometer-based fall detection method using a self-constructing cascade-AdaBoost-SVM classifier. *IEEE J. Biomed. Health Inf.* 17 (2).

Chen, H., Chandrasekar, V., Cifelli, R., & Xie, P. 2019. A machine learning system for precipitation estimation using satellite and ground radar network observations. *IEEE Transactions on Geoscience and Remote Sensing*, 58(2), 982-994.

Cornuéjols.A., Miclet.L. “ Apprentissage Artificiel, concepts et algorithmes” 2ieme Edition. ISBN: 978-2-212-12471-2, Eyrolles , 2010.

Cortes, C. and Vapnik, V., 1995. Support-Vector Networks. *Machine Learning*, 20, 273-297. <http://dx.doi.org/10.1007/BF00994018>

Cover, T., & Hart, P., 1967. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1), 21-27. doi:10.1109/tit.1967.1053964

Crammer, K., Singer, Y., 2002. On the learnability and design of output codes for multiclass problems. *Mach. Learn.* 47 (2–3), 201–233.

Bibliographie

Dietterich, T.G., 2000. Ensemble Methods in Machine Learning. In: Multiple Classifier Systems. MCS 2000. Lecture Notes in Computer Science, vol. 1857. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-45014-9_1. 2000.

Du, P., Bai, X., Tan, K., et al., 2020. Advances of four machine learning methods for spatial data handling: a review. *J. Geovis. Spat. Anal.* 4,13. <https://doi.org/10.1007/s41651-020-00048-5>.

EUMETSAT, 2004. Applications of Meteosat Second Generation—conversion from counts to radiances and from radiances to brightness temperatures and reflectance http://oiswww.eumetsat.org/WEBOPS/msg_interpretation/index.html.

Eumetsat, 2005. MSG : Meteosat Second Génération En Orbit- En Service.

http://www.eumetsat.int/groups/cps/documents/document/pdf_msg02_fr.pdf.

EUMETSAT, 2006. MSG CHANNELS Interpretation Guide: Weather, Surface conditions and Atmospheric constituents, <http://oiswww.eumetsat.org>.

Eumetsat, 2013. MSG Level 1.5 Image Data Format Description.

<http://www.eumetsat.int/website/home/Data/Products/Formats/index.html>

Feidas H., Kokolatos G., Negri A., Manyin M., Chrysoulakis N. and Kamarianakis Y. 2008: Validation of an infrared-based satellite algorithm to estimate accumulated rainfall over the Mediterranean basin, *Theor. Appl. Climatol.* DOI 10.1007/s00704-007-0360-y.

Feidas H. and Giannakos A. 2010: Identifying precipitating clouds in Greece using multispectral infrared Meteosat Second Generation satellite data. *Theor. Appl. Climatol.* doi:10.1007/s00704-010-0316-5.

Feidas, H., Giannakos, A., 2011. Classifying convective and stratiform rain using multispectral infrared Meteosat Second Generation satellite data. *Theor. Appl. Climatol.* <https://doi.org/10.1007/s00704-011-0557-y>.

Freund, Y., Schapire, R.E., 1997. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *J. Comput. Syst. Sci.* 55 (1), 119–139.

Bibliographie

Galar, M., Fernández, A., Barrenechea, E., Bustince, H., & Herrera, F. 2011. An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes. *Pattern Recognition*, 44(8), 1761-1776.

Goodfellow, I., Bengio, Y., & Courville, A. 2016. *Deep learning*. MIT press.

Griffith C. G., Woodley W. L., Grube P. G., Martin D. W., Stout J. and Sikdar D. N. 1978: Rain estimation from geosynchronous satellite imagery - visible and infrared studies. *Monthly Weather Review*, Vol 106, N°8, pp 1153-1171.

Gupta T.K., Raza K., 2019. Chapter 7 - Optimization of ANN Architecture: A Review on Nature-Inspired Techniques, Editor(s): NilanjanDey, SurekhaBorra, Amira S. Ashour, Fuqian Shi, Machine Learning in Bio-Signal Analysis and Diagnostic Imaging, Academic Press, 2019, Pages 159–182.

Han W., Stamnes K. and Lubin D. 1999: Remote sensing of surface and cloud properties in the Arctic from AVHRR measurements. *J. Appl. Meteorol.*, 38, 989-1012.

Hinton, G. E., Osindero, S., & Teh, Y. W., 2006. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7), 1527-1554.

Houze, R.A, 1997: Stratiform Precipitation in Regions of Convection: A Meteorological Paradox?, *Bulletin of the American Meteorological Society.*, **78**, 2179-2196.

Houze Jr. 1993: Cloud dynamics. *Academic Press*, 573pp

Inoue, T. (1985). The use of the radiative transfer equation in the study of the remote sensing of cloud parameters. *Journal of the Atmospheric Sciences*, 42(6), 601-612.

Jensen, J.R., 2005. "Introductory Digital Image Processing: A Remote Sensing Perspective". 3rd Edition, Prentice Hall, Upper Saddle River, 505-512.

Jones, D. 1956. Rainfall drop size-distribution and radar reflectivity. *ISWS Contract Report CR 009*.

Kanchan, P., & Shardoor, N. K. B., 2021. Rainfall analysis and forecasting using deep learning technique. *Journal of Informatics Electrical and Electronics Engineering (JIEEE)*, 2(2), 1-11.

Bibliographie

Kavitha, M., Gayathri, R., Polat, K., Alhudhaif, A., Alenezi, F., 2022. Performance evaluation of deep e-CNN with integrated spatial-spectral features in hyperspectral image classification. *Measurement* 191, 110760.

King, M. D., S.-C. Tsay, S. E. Platnick, M. Wang, and K.-N. Liou, 1997. Cloud retrieval algorithms for MODIS: Optical thickness, effective particle radius, and thermodynamic phase. MODIS Algorithm Theoretical Basis Document ATBD-MOD-05, NASA, 78 pp.

Kokhanovsky, A. A., & Nauss, T. 2006. Reflection and transmission of solar light by clouds: asymptotic theory. *Atmospheric Chemistry and Physics*, 6(12), 5537-5545.

Kühnlein, M ; Appelhans, T; Thies, B; Nauß, T. 2014a. Precipitation Estimates from MSG SEVIRI Daytime, Nighttime, and Twilight Data with Random Forests. *Journal of Applied Meteorology and Climatology*, 53(11), 2457–2480. doi:10.1175/jamc-d-14-0082.1

Kuhnlein, M., Appelhans, T., Thies, B., Nauss, T., 2014b. Improving the accuracy of rainfall rates from optical satellite sensors with machine learning — a random forests-based approach applied to MSG SEVIRI. *Remote Sens. Environ.* 141, 129–143.

Kurino T. 1997: A satellite infrared technique for estimating deep/shallow convective and stratiform precipitation. *Adv. Space Res.*, 19, 511-514.

Lacaze B. and Bergès J.C. 2005: Contribution of Meteosat second generation (MSG) to drought early warning. *Proceedings of the International Conference: Remote sensing and geoinformation processing in the assessment and monitoring of land degradation and desertification*, pp 1-8.

Lazri M., Ouallouche F., Ameur S., Brucker J. M. and Mohia Y. 2012: Identifying Convective and Stratiform Rain by Confronting SEVIRI Sensor Multispectral Infrared to Radar Sensor Data Using Neural Network, *Sensors & Transducers Journal*, Vol. 145, Issue 10, pp. 19-32.

Lazri M., 2013. Analyse des variations spatio-temporelles des précipitations par télédétection active et passive. Thèse doctorale. Université Mouloud Mammeri, Tizi-Ouzou, Algérie.

Lazri, M., Ameur, Z., Ameur, S., Mohia, Y., Brucker, J. M., & Testud, J. (2013a). Rainfall estimation over a Mediterranean region using a method based on various spectral parameters of SEVIRI-MSG. *Advances in space research*, 52(8), 1450-1466.

Bibliographie

Lazri M., Ameer S., Brucker J. M., Testud J., Hamadache B., Hameg S., Ouallouche F. and Mohia Y. 2013b: Identification of raining clouds using a method based on optical and microphysical cloud properties from Meteosat second generation daytime and nighttime data, *Appl Water Sci*, DOI 10.1007/s13201-013-0079-0.

Lazri, M., Ameer, S., & Mohia, Y. (2014a). Instantaneous rainfall estimation using neural network from multispectral observations of SEVIRI radiometer and its application in estimation of daily and monthly rainfall. *Advances in Space Research*, 53(1), 138-155.

Lazri, M., Ameer, S., Brucker, J.M., Ouallouche, F., 2014b. Convective rainfall estimation from MSG/SEVIRI data based on different development phase duration of convective systems (growth phase and decay phase). *Atmos. Res.* 147–148, 38–50.

Lazri, M., & Ameer, S., 2016. Using cloud water path and cloud top temperature for estimating convective and stratiform rainfall from SEVIRI daytime data. *Arabian Journal of Geosciences*, 9(11), 577.

Lazri, M., & Ameer, S., 2018. Combination of support vector machine, artificial neural network and random forest for improving the classification of convective and stratiform rain using spectral features of SEVIRI data. *Atmospheric research*, 203, 118-129. <https://doi.org/10.1016/j.atmosres.2017.12.006>

Lazri, M., Labadi, K., Brucker, J. M., & Ameer, S., 2020. Improving satellite rainfall estimation from MSG data in Northern Algeria by using a multi-classifier model based on machine learning. *Journal of Hydrology*, 584, 124705.

LeCun, Y., 1989. Generalization and network design strategies. *Connectionism in perspective*, 19, pp.143-155

LeCun, Y., Bengio, Y. and Hinton, G., 2015. Deep learning. *nature*, 521(7553), pp.436-444.

Lensky I.M., Rosenfeld D (2003a). A night-rain delineation algorithm for infrared satellite data based on microphysical considerations. *J Appl Meteorol* 42:1218–1226

Lensky, I.M., Rosenfeld, D., 2003b. Satellite-based insights into precipitation formation processes in continental and maritime convective clouds at nighttime. *J. Appl. Meteorol.* 42, 1227–1233.

Bibliographie

Levizzani V., Schmetz J., Lutz H. J., Kerkmann J., Alberoni P. P., Cervino M. 2001: Precipitation estimations from geostationary orbit and prospects for Meteosat Second Generation, *Meteorological Applications*, 8, 23–41.

Levizzani, V., 2003. Satellite rainfall estimations: new perspectives for meteorology and climate from the EURAINSAT project. *Ann. Geophys.* 46, 363–372.

Li, X., Wang, L., Sung, E. (2005). A study of AdaBoost with SVM based weak learners. In: *Proceedings. 2005 IEEE International Joint Conference on Neural Networks*, vol. 1, pp. 196–201. doi: 10.1109/IJCNN.2005.1555829.

Lorena, A. C., De Carvalho, A. C., & Gama, J. M. (2008). A review on the combination of binary classifiers in multiclass problems. *Artificial Intelligence Review*, 30, 19–37.

Mace, G.G., T.P Ackerman, P. Minnis and D.F. Young, 1998a: Cirrus layer microphysical properties derived from surface-based millimetre radar and infrared interferometer, *J. Geophys. Res.*, 103, D18, 23207–23216.

Maia-Bresson, A. (1996). Estimation des précipitations par satellite: analyse et application d'une méthode combinée Infrarouge & micro-ondes. (Thèse de doctorat). Université Paris VII, Paris.

Marshall, J. S., & Palmer, W. M. K. (1948). The distribution of raindrops with size. *Journal of Atmospheric Sciences*, 5(4), 165–166.

Matrosov, S.Y., T. Uttal, J.B. Snider and R.A. Kropfli, 1992: Estimation of ice clouds parameters from ground-based infrared radiometer and radar measurements, *J. Geophys. Res.*, 97, 11567–11574.

Maxwell, A. E., Warner, T. A., & Fang, F. (2018). Implementation of machine-learning classification in remote sensing: An applied review. *International journal of remote sensing*, 39(9), 2784–2817.

Mehta, P., Bukov, M., Wang, C. H., Day, A. G., Richardson, C., Fisher, C. K., & Schwab, D. J., 2019. A high-bias, low-variance introduction to machine learning for physicists. *Physics reports*, 810, 1–124.

Bibliographie

Milford, J.R. and G. Dugdale, 1989: Estimation of rainfall using geostationary satellite data. *Applications of Remote Sensing in Agriculture*, Proc. of 48th Easter School in Agricultural Science, University of Nottingham, Butterworth, London.

Milford, J. R., & Dugdale, G. (1990). Estimation of rainfall using geostationary satellite data. *Applications of remote sensing in agriculture*, 97, 110.

Mohia, Y., Ameer, S., Lazri, M., & Brucker, J. M., 2017. Combination of spectral and textural features in the MSG satellite remote sensing images for classifying rainy area into Different classes. *Journal of the Indian Society of Remote Sensing*, 45(5), 759-771.

Mohia, Y., Absi, R., Lazri, M., Labadi, K., Ouallouche, F., & Ameer, S., 2023. Quantitative Estimation of Rainfall from Remote Sensing Data Using Machine Learning Regression Models. *Hydrology*, 10(2), 52.

Morau, A., Dewitte, S., Cornelis, B., & Munteanu, A. 2019. Deep learning for precipitation estimation from satellite and rain gauges measurements. *Remote Sensing*, 11(21), 2463.

Nakajima T. and M.D. King, 1990: Determination of the optical thickness and effective particle radius of clouds from reflected solar radiation measurements, Part I: Theory, *J. Atmos. Sci.*, 47, 15, 1878-1893

Nauss, T., Kokhanovsky, A. A., Nakajima, T. Y., Reudenbach, C., & Bendix, J., 2005. The intercomparison of selected cloud retrieval algorithms. *Atmospheric Research*, 78(1-2), 46-78.

Nauss T. and Kokhanovsky A. A. 2006: Discriminating raining from non-raining clouds at mid latitudes using multispectral satellite data. *Atmos. Chem. Phys.*, 6, 5031–5036.

Negri, A.J., Adler, R.F. and Wetzel, P.J., 1984, Satellite rain estimation: An analysis of the Griffith-Woodley technique. *Journal of Climate and Applied Meteorology*, 23, pp.102–116.

Negri, A.J., Adler, R.F., 1993. An intercomparison of three satellite infrared rainfall techniques over Japan and surrounding waters. *J. Appl. Meteorol.* 32, 357–373.

Opitz, D., Maclin, R., 1999. Popular ensemble methods: an empirical study. *J. Artif. Intell. Res.* 11 (1), 169–198.

Bibliographie

- Ortiz-García, E. G., Salcedo-Sanz, S., & Casanova-Mateo, C., 2014. Accurate precipitation prediction with support vector classifiers: A study including novel predictive variables and observational data. *Atmospheric research*, 139, 128-136. doi:10.1016/j.atmosres.2014.01.012
- Ouallouche, F., Lazri, M., Ameer, S., 2018. Improvement of rainfall estimation from MSG data using Random Forests classification and regression. *Atmos. Res.* 211, 62–72.
- Oukali, S., Lazri, M., Labadi, K., Brucker, J.M., Ameer, S., 2019. Development of a hybrid classification technique based on deep learning applied to MSG/SEVIRI multispectral data. *J. Atmos. Sol.-Terrest. Phys.* 193,105062.
- Parol F., Buriez J. C., Brogniez G. and Fouquart Y. 1991: Information content of AVHRR channels 4 and 5 with respect to the effective radius of cirrus cloud particles. *J. Appl. Meteorol.*, 30, 973-984.
- Petty G. W. 1994: Physical retrievals of over-ocean rain rate from multichannel microwave imaging. Part I: Theoretical characteristics of normalized polarization and scattering indices. *Meteor. Atmos. Phys.*, 54, 79–100
- Phung, ; Rhee, (2019). *A High-Accuracy Model Average Ensemble of Convolutional Neural Networks for Classification of Cloud Image Patches on Small Datasets*. *Applied Sciences*, 9(21), 4500–. doi:10.3390/app9214500
- Pollack, J.B., 1990. Recursive distributed representations. *Artificial Intelligence*, 46(1-2), pp.77-105.
- Prasad, N., Reddy, P. K., & Naidu, M. M., 2013. A novel decision tree approach for the prediction of precipitation using entropy in sliq. In *2013 UKSim 15th International Conference on Computer Modelling and Simulation* (pp. 209-217). IEEE. doi: 10.1109/UKSim.2013.74.
- Refaeilzadeh, P., Tang, L., Liu, H. (2009). Cross-Validation. In: LIU, L., ÖZSU, M.T. (eds) *Encyclopedia of Database Systems*. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-39940-9_565

Bibliographie

Reudenbach, C., Heinemann, G., Heuel, E., Bendix, J., & Winiger, M. 2001. Investigation of summertime convective rainfall in Western Europe based on a synergy of remote sensing data and numerical models. *Meteorology and Atmospheric Physics*, 76, 23-41.

Reudenbach, C., 2003. Konvektive Sommerniederschläge in Mitteleuropa. Eine Kombination aus Satelliten fernerkundung und numerischer Modellierung zur automatischen Erfassung mesoskaliger Niederschlagsfelder. *Bonner Geographische Abhandlungen* 109, 152.

Rivrain J.-C. 1997: Les épisodes orageux à précipitations extrêmes sur les régions méditerranéennes de la France, *Phénomènes remarquables no. 4, Météo-France, Toulouse, (in French)*.

Roebeling, R.A., Feijt, A.J., Stammes, P., 2006. Cloud property retrievals for climate monitoring: implications of differences between SEVIRI on METEOSAT-8 and AVHRR on NOAA-17. *J. Geophys. Res.* 11, D20210. <https://doi.org/10.1029/2005JD006990>.

Roebeling R. A. and Holleman I. 2009: Validation of rain rate retrievals from SEVIRI using weather radar observations, *J. Geophys. Res.* 114, D21202, doi:10.1029/2009JD012102.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6), 386.

Rossow, W. B., & Schiffer, R. A., 1999. Advances in understanding clouds from ISCCP. *Bulletin of the American Meteorological Society*, 80(11), 2261-2288.

Rumelhart, D.E., Hinton, G.E. and Williams, R.J., 1985. Learning internal representations by error propagation (No. ICS-8506). California Univ San Diego La Jolla Inst for Cognitive Science.

Sadeghi, M., Asanjan, A. A., Faridzad, M., Nguyen, P. H. U., Hsu, K., Sorooshian, S., & Braithwaite, D. A. N. 2019. PERSIANN-CNN: Precipitation estimation from remotely sensed information using artificial neural networks—convolutional neural networks. *Journal of Hydrometeorology*, 20(12), 2273-2289.

Bibliographie

Samuel, Arthur L, 1959. Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 3(3):210–229.

Sauvageot, H., 1992. *Radar Meteorology*, Artech House, Boston, 361.

Sauvageot, H., 1994 : Rainfall measurement by radar : A review , *Atmospheric Research* 35 (1), 27-54. doi:10.1016/0169-8095(94)90071-x

Schapire, R. E. (1990). The strength of weak learnability. *Machine learning*, 5, 197-227.

Schapire, R.E., Singer, Y., 1999. Improved boosting algorithms using confidence-rated predictions. *Mach. Learn.* 37, 297–336. <https://doi.org/10.1023/A:1007614523901>.

Seck I., 2020 .Estimation et Prévision immédiate des précipitations sur un bassin urbain. Thèse de doctorat, spécialité : Physique-chimie de l'atmosphère et climat. Ecole Doctorale des Sciences Fondamentale, Université Clermont Auvergne, France.

Sehad, M., Lazri, M., & Ameer, S., 2017. Novel SVM-based technique to improve rainfall estimation over the Mediterranean region (north of Algeria) using the multispectral MSG SEVIRI imagery. *Advances in Space Research*, 59(5), 1381-1394.

Sehad, M., Ameer, S., 2020. A multilayer perceptron and multiclass support vector machine based high accuracy technique for daily rainfall estimation from MSG SEVIRI data. *Adv. Space Res.* 65 (4), 1250–1262.

Shtanko, A., Kulik, S., 2021. Increasing the effectiveness of intelligent module by enlarging training dataset from real data. *Procedia Comput. Sci.* 190, 712–716.

Smith, E. A. and Coauthors, 1998 : Results of wetnet pip-2 project. *J. Atmos. Sci.*, 55, 1483–1536.

Sorooshian, S., Hsu, K. L., Gao, X., Gupta, H. V., Imam, B. & Braithwaite, D. 2000. Evaluation of PERSIANN system satellite-based estimates of tropical rainfall. *Bull. Am. Meteorol. Soc.* 81: 2035–2046. DOI:[10.1175/1520-0477\(2000\)081<2035:EOPSSE>2.3.CO;2](https://doi.org/10.1175/1520-0477(2000)081<2035:EOPSSE>2.3.CO;2)

Spencer, R. W., Goodman, H. M., & Hood, R. E., 1989. Precipitation retrieval over land and ocean with the SSM/I: Identification and characteristics of the scattering signal. *Journal of Atmospheric and Oceanic Technology*, 6(2), 254-273.

Bibliographie

- Strabala, K. I., Ackerman, S. A., & Menzel, W. P. 1994. Cloud Properties inferred from 8 12- μ m Data. *Journal of Applied Meteorology and Climatology*, 33(2), 212-229.
- Takbiri, Z., Ebtehaj, A., Foufoula-Georgiou, E., Kirstetter, P. E., & Turk, F. J. 2019. A Nested K-Nearest Prognostic Approach for Microwave Precipitation Phase Detection over Snow Cover. *arXiv preprint arXiv:1902.09578*.
- Tapiador, F. J., Kidd, C., Levizzani, V. & Marzano, F. S. 2004. A neural networks-based fusion technique to estimate half hourly rainfall estimates at 0.1° resolution from satellite passive microwave and infrared data. *J. Appl. Meteorol.*43, 309, pp. 98–116.
- Tebbi, M. A., & Haddad, B., 2016. Artificial intelligence systems for rainy areas detection and convective cells' delineation for the south shore of Mediterranean Sea during day and nighttime using MSG satellite images. *Atmospheric research*, 178, 380-392.
- Thies, B., Nauss, T. and Bendix J., 2008 a. Discriminating raining from non-raining clouds at mid-latitudes using meteosat second generation daytime data. *Atmos. Chem. Phys.*, 8, 2341–2349.
- Thies, B., Nauss T. and Bendix J. 2008b. Delineation of raining from nonraining clouds during nighttime using Meteosat-8 data. *Meteorol Appl* 15:219–230.
- Thies, B., Nauss, T., & Bendix, J. 2008c. Precipitation process and rainfall intensity differentiation using Meteosat second generation spinning enhanced visible and infrared imager data. *Journal of Geophysical Research: Atmospheres*, 113(D23).
- Thillet, J.J. 1997: *La météo de montagne. Les guides du club alpin français. Seuil*
- Tian, W., Yi, L., Liu, W., Huang, W., Ma, G., & Zhang, Y. 2020. Ground radar precipitation estimation with deep learning approaches in meteorological private cloud. *Journal of cloud computing*, 9(1), 22.
- Tjemkes S. A., Van De Berg L. and Schmetz J. 1997: Warm water vapour pixels over high clouds as observed by Meteosat. *Contributions to Atmospheric Physics*, 70: 15-21.
- Trigo, R. M., Xoplaki, E., Zorita, E. et al. 2006: Relationship between variability in the Mediterranean region and mid-latitude variability, *Mediterranean Climate Variability*, Elsevier, B. V., 179–226, 2006.

Bibliographie

Tsay, J. D., Kao, K., Chao, C. C., & Chang, Y. C., 2020. Deep learning for satellite rainfall retrieval using Himawari-8 multiple spectral channels. *Preprints* **2020**, 2020100648.

<https://doi.org/10.20944/preprints202010.0648.v1>

Turini, N., Thies, B., & Bendix, J. 2019. Estimating high spatio-temporal resolution rainfall from MSG1 and GPM IMERG based on machine learning: Case study of Iran. *Remote Sensing*, *11*(19), 2307.

Turini, N., Thies, B., Horna, N., & Bendix, J. 2021. Random forest-based rainfall retrieval for Ecuador using GOES-16 and IMERG-V06 data. *European Journal of Remote Sensing*, *54*(1), 117-139.

Waymire E, V Gupta, I Rodriguez-Iturbe (1984) A spectral theory of rainfall intensity at the meso-beta scale, *Water Resources Research*, *20* (10):1453–1465.

Wickramaratna, J., Holden, S., Buxton, B., 2001. Performance degradation in boosting. In: Kittler, J., Roli, F. (Eds.), *Proceedings of the 2nd International Workshop on Multiple Classifier Systems MCS2001*. Volume 2096 of LNCS. Springer, pp. 11–21.

Wilheit, T. T., R. Adler, S. Avery, E. Barrett, P. Bauer, W. Berg, A. Chang, J. Ferriday, N. Grody, S. Goodman, C. Kidd, D. Kniveton, C. Kummerow, A. Mugnai, W. Olson, G. Petty, A. Shibata, E. Smith, and R. Spencer, 1994 : Algorithms for the retrieval of rainfall from passive microwave measurements. *Remote Sens. Rev.*, *11*, 163–194.

Wolff, C., Vaillant, P., Radartutorial, chapitre 2A : Radar météorologique, version 19 septembre 2011.

Wolters, E.L.A., Roebeling, R.A., Feijt, A.J. 2008. Evaluation of cloud phase retrieval methods for SEVIRI onboard Meteosat-8 using groundbased lidar and cloud radar data. *J. Appl. Meteorol. Clim.* *47*, 1723– 1738, <http://dx.doi.org/10.1175/2007JAMC1591.1>.

Résumé

Dans ce travail nous implémentons l'algorithme AdaBoost pour optimiser les résultats des classifications d'intensités de précipitations réalisées par la stratégie un-contre-tous utilisant les vecteurs machines de support (OvA-SVM pour One-vs-All-SVM) au nord de l'Algérie. Le modèle développé qui combine l'algorithme AdaBoost avec un SVM multiclassés, est appliqué aux images du satellite MSG (Meteosat Second Generation), et qui est appris et validé par une comparaison avec les données Radar correspondantes. D'autres variantes pour construire des SVM multiclassés, tels que l'OvO-SVM (One versus One SVM), SBT-SVM (Slant Binary Tree SVM) et DDAG-SVM (Decision Directed Acyclic Graph) sont également implémentés sur lesquelles nous avons testé l'algorithme AdaBoost. L'étude a montré que l'algorithme AdaBoost a mieux fonctionné et a donné de meilleurs résultats dans le cas de la variante OvA-SVM par rapport aux autres variantes.

Afin d'évaluer le modèle élaboré, certaines techniques de classification, telles que l'ECST Enhanced Convective Stratiform Technique (ECST), le SART où sont combinés la machine à vecteurs de support avec le réseau de neurones artificiels et les classificateurs de forêt aléatoire, la technique de délimitation des zones de pluie convective/stratiforme (CS-RADT) et la technique de forêt aléatoire (RFT) sont appliquées.

Les résultats de classification obtenus montrent que la méthode ensembliste AdaOvA-SVM présente des performances très intéressantes et satisfaisantes avec des améliorations des précisions de classification et estimation des précipitations par rapport aux autres méthodes comparées utilisées précédemment.

Mots clés : AdaBoost ; SVM ; Satellite MSG ; classification ; apprentissage automatique ; précipitations.

Abstract:

In this work, we implement the AdaBoost algorithm to optimize the classifications results of precipitations intensities carried out by One-versus-All strategy using Support Vector Machine (OvA-SVM) in Northern Algeria. The model developed which combines the AdaBoost algorithm with a multiclass SVM is applied to images from the MSG (Meteosat Second Generation) satellite, and it is learned and validated by comparison with the corresponding Radar data. Other variants to build multiclass SVMs, such as the OvO-SVM (One versus One SVM), SBT-SVM (Slant Binary Tree SVM) and DDAG-SVM (Decision Directed Acyclic Graph) are also implemented on which we tested the AdaBoost algorithm. The study showed that the AdaBoost algorithm performed better in the case of the OvA-SVM variant compared to the other variants.

In order to evaluate the elaborated model, some classification techniques, such as the Enhanced Convective Stratiform Technique (ECST), the SART where the Support vector machine, Artificial neural network and Random forest classifiers are combined, the Convective/Stratiform Rain Area Delineation Technique (CS-RADT) and the Random Forest technique (RFT) are applied.

The classification results obtained show that the ensemble method AdaOvA-SVM presents very interesting and satisfying performances with improvements in classification accuracies and precipitation estimation compared to the other methods previously used.

Keywords: AdaBoost; SVM; MSG Satellite; Classification; Machine Learning; Precipitations.