UNIVERSITÉ DE NGAOUNDÉRÉ

FACULTÉ DES SCIENCES

DÉPARTEMENT DE MATHEMATIQUES ET INFORMATIQUE

THE UNIVERSITY OF NGAOUNDERE

FACULTY OF SCIENCE

DEPARTMENT OF MATHEMATICS AND COMPUTER SCIENCE

ÉCOLE DOCTORALE : SCIENCES, TECHNOLOGIE ET INGÉNIERIE (STI)

UFD : MATHÉMATIQUES, INFORMATIQUE, INGÉNIERIE ET APPLICATION (M2IAP)

Deep Analysis and prediction of chikungunya using ensemble regression approach

Mémoire en vue d'obtention du Master II

Filière : Informatique

Spécialité : Systèmes et Logiciels en Environnements distribués

Par

MOHAMED EL BACHIR BOUBA NGANADAKOUA

(Master I Systèmes et Logiciels en Environnements distribués)

Matricule : 19A666FS

Sous la direction de :

Pr. Dr. Ing Dayang Paul
Maître de Conférence
Faculté des Sciences
Université de Ngaoundéré

Dr. ABBOUBAKAR Hamadjam
Chargé de Cours
Institut Universitaire de Technologie
Université de Ngaoundéré

Année académique 2023-2024

Dédicace

à mon père BOUBA Nana Dekwa,
à ma mère HABIBA Kouyessi,
à mes sœurs.

Remerciements

La réalisation de ce mémoire a été un long voyage, parsemé de défis et d'apprentissages, et n'aurait pas été possible sans le soutien et l'assistance de nombreuses personnes. C'est avec une profonde gratitude que je tiens à exprimer mes sincères remerciements à tous ceux qui ont contribué, de près ou de loin, à la concrétisation de ce travail.

- Tout d'abord, je remercie Allah le Tout-Puissant qui m'a donné la volonté et la force nécessaires pour parfaire ce travail et le mener à terme ;
- Le Doyen de la Faculté des Sciences de l'Université de Ngaoundéré, le Professeur NGAMENI Emmanuel, pour son soutien et ses encouragements ;
- Le Chef du Département de Mathématiques et Informatique de la Faculté des Sciences de l'Université de Ngaoundéré : le Pr. Dr. Ing DAYANG Paul pour le suivi de notre formation et la supervision de ce travail ;
- Mon directeur de mémoire, Dr. ABBOUBAKAR Hamadjam, Chargé de Cours à l'Université de Ngaoundéré, au Département de Génie Informatique à l'Institut Universitaire de Technologie, pour m'avoir proposé ce sujet et pour ses précieux conseils ;
- Les enseignants du Département de Mathématiques et Informatique de la Faculté des Sciences de l'Université de Ngaoundéré, ainsi que les membres du jury, pour avoir accepté d'examiner ce travail ;
- Ma famille, pour leurs nombreux conseils, encouragements et soutien tout au long de mes études ;
- Mes camarades de promotion, pour la convivialité, le partage de connaissances et l'entraide dont nous avons bénéficié ensemble.

À tous, un grand merci.

Table des matières

Dédicace **i**

Remerciements **ii**

Table des matières **iii**

Liste de figures **vi**

Liste des tableaux **vii**

Notation et abréviations **viii**

Résumé **ix**

Abstract **x**

Introduction Générale **0**

1 Épidémiologie de la Chikungunya **1**

1.1 Origine de la Chikungunya 1

1.2 Agent Pathogène 2

1.3 Mode de Transmission 2

1.3.1 Vecteurs : Les moustiques Aedes 3

1.3.2 La transmission 3

1.4 Symptômes et Diagnostic 4

1.4.1 Symptômes 4

1.4.2 Méthodes de diagnostic 4

1.5 Méthodes de Contrôle et Traitement 4

1.5.1 Méthodes de contrôle 4

1.5.2 Options de traitement 5

1.6 Cas du Tchad 5

1.6.1 Point saillants 6

1.6.2 Contexte 6

1.7 Cas du Brésil 7

1.7.1 Point saillants 7

1.7.2 Contexte 7

1.8 Cas du Paraguay . . . 8
1.8.1 Point saillants . . . 8
1.8.2 Contexte . . . 9

2 Revue de la Littérature et Concepts de Base 10
2.1 Apprentissage Automatique . . . 10
2.1.1 Techniques d'Apprentissage Automatique . . . 11
2.1.1.1 Apprentissage Supervisé . . . 11
2.1.1.2 Apprentissage Non Supervisé . . . 12
2.1.1.3 Apprentissage Semi-Supervisé . . . 12
2.2 Introduction au Deep Learning . . . 12
2.2.1 Les bases du Deep Learning . . . 12
2.2.1.1 Réseaux de neurones . . . 12
2.2.1.2 Fonctionnement du réseau . . . 13
2.2.2 Fonctions d'activation . . . 14
2.2.3 Entraînement des modèles . . . 14
2.2.3.1 Fonction de coût . . . 14
2.2.3.2 Optimisation . . . 15
2.2.4 Applications du Deep Learning . . . 15
2.3 Ensemble Learning ou Apprentissage par Ensemble . . . 16
2.3.1 Définition . . . 16
2.3.2 principes de fonctionnement . . . 16
2.3.3 Techniques d'Ensemble Learning . . . 17
2.3.3.1 Échantillonnage et sélection des données : Diversité . . . 17
2.3.3.2 Formation des classificateurs de membres . . . 17
2.3.3.3 Combinaison des membres de l'ensemble . . . 18
2.3.4 Algorithmes populaires basés sur l'ensemble learning . . . 19
2.3.4.1 Bagging . . . 19
2.3.4.2 Boosting and AdaBoost . . . 20

3 Implémentation des Modèles 22
3.1 Introduction . . . 22
3.2 Présentation des Outils Utilisés . . . 22
3.2.1 Langages et librairies . . . 22
3.2.2 Infrastructure matérielle . . . 22
3.3 Méthodes de l'Apprentissage Automatique . . . 22
3.3.1 Le jeu de données . . . 22
3.3.2 Préparation et nettoyage des données . . . 22
3.3.3 Implémentation des modèles de machine learning . . . 22
3.4 Méthodes de l'Apprentissage Profond . . . 22
3.4.1 Le jeu de données . . . 23

3.4.2 Préparation des données pour le deep learning 23
3.4.3 Implémentation des modèles de deep learning 23

General Conclusion **25**

Bibliographie **26**

Liste de figures

1.1 CHIKV in the Western Hemisphere. 2
1.2 Electron microscopic image of chikungunya virus 2
1.3 Aedes aegypti mosquito full of blood 3
1.4 Transmission mode Chikungunya 3
1.5 Region infecté au Tchad 6
1.6 Evolution journalière des cas et décès du Chikungunya 6
1.7 Chikungunya Cases brazil 7
1.8 Map of the epicenters of chikungunya cases 8
1.9 Expansion of the chikungunya East/Central/South/African lineage epidemic in Paraguay 9
1.10 Weekly reported chikungunya cases 9

2.1 Machine Learning Tree 11
2.2 Flux de travail de l'apprentissage supervisé 11
2.3 Apprentissage Non Supervisé 12
2.4 Neurone Biolique en Artificiel 13
2.5 Foward-Backward-Propagation 13
2.6 Réduction de la variabilité à l'aide de systèmes d'ensemble 16

Liste des tableaux

Notation et abréviations

CHIKV	Chikungunya virus
WHO	World Health Organization
PAHO	Pan American Health Organization
OMS	Organisation Mondiale de la Sante
CDC	Centers for Disease Control and Prevention
Ae	Aedes
CHIKVD	Chikungunya Virus Diseases
ML	Machine learning
ReLu	Rectified linear unit
RNA	Réseau de Neurone Artificiel
RNN	Recurrent neural network
RMSE	Root Squared Mean Error
CNN	Convolution neural network

Résumé

Le virus du chikungunya, transmis principalement par les moustiques **Aedes aegypti** et **Aedes albopictus**, représente une menace croissante pour la santé publique mondiale en raison de sa propagation rapide et de ses effets débilitants. Récemment, des épidémies ont été signalées non seulement en Afrique centrale et orientale, mais aussi en Amérique du Sud et en Asie du Sud-Est. Prédire ces épidémies reste un défi majeur en raison de l'interaction complexe entre les facteurs environnementaux, climatiques et biologiques. Les approches traditionnelles de surveillance épidémiologique se révèlent souvent insuffisantes pour anticiper les épidémies de manière proactive.

Cette recherche se concentre sur le **Tchad**, le **Brésil**, et le **Paraguay**, trois pays où le chikungunya a émergé comme un problème significatif de santé publique. En raison de la disponibilité limitée de données au Tchad, des données supplémentaires provenant du Brésil et du Paraguay sont intégrées pour renforcer l'analyse. Le but de cette thèse est de développer et d'évaluer des modèles prédictifs pour les épidémies de chikungunya en utilisant des techniques avancées d'apprentissage automatique, en particulier la régression d'ensemble.

Les modèles choisis pour cette étude incluent le **Random Forest Regressor**, le **XGBoost Regressor** optimisé via **Grid Search**, ainsi qu'un modèle d'ensemble (**Voting Regressor**) combinant **Linear Regression**, **Decision Tree Regressor**, et le **XGBoost Regressor** optimisé. Ces modèles seront formés et validés à partir de données épidémiologiques et climatiques.

Cette thèse commence par une analyse approfondie de l'épidémiologie du chikungunya et de ses dynamiques de transmission, suivie d'une revue des applications actuelles de l'intelligence artificielle dans la prédiction des maladies. Les modèles sont ensuite testés et évalués sur la base de critères de performance tels que la sensibilité, la spécificité, et la valeur prédictive.

Les résultats de cette recherche montrent l'efficacité des techniques de régression d'ensemble pour améliorer la précision des prévisions épidémiologiques, et offrent des perspectives nouvelles pour l'intervention en santé publique dans les régions affectées.

Mots-clés : *Chikungunya, Régression d'ensemble, Random Forest, XGBoost, Apprentissage Automatique, Données Climatiques, Prévision Épidémiologique*

Abstract

The chikungunya virus, transmitted mainly by the mosquitoes **Aedes aegypti** and **Aedes albopictus**, represents a growing threat to global public health due to its rapid spread and debilitating effects. Recently, epidemics have been reported not only in Central and East Africa, but also in South America and Southeast Asia. Predicting these epidemics remains a major challenge due to the complex interplay between environmental, climatic and biological factors. Traditional epidemiological surveillance approaches often prove insufficient to proactively anticipate epidemics.

This research focuses on **Chad**, **Brazil**, and **Paraguay**, three countries where chikungunya has emerged as a significant public health problem. Due to limited data availability in Chad, additional data from Brazil and Paraguay are incorporated to strengthen the analysis. The aim of this thesis is to develop and evaluate predictive models for chikungunya epidemics using advanced machine learning techniques, in particular ensemble regression.

The models chosen for this study include the **Random Forest Regressor**, the **XGBoost Regressor** optimized via **Grid Search**, as well as an ensemble model (**Voting Regressor**) combining **Linear Regression**, **Decision Tree Regressor**, and the optimized **XGBoost Regressor**. These models will be trained and validated using epidemiological and climatic data.

This thesis begins with an in-depth analysis of the epidemiology of chikungunya and its transmission dynamics, followed by a review of current applications of artificial intelligence in disease prediction. The models are then tested and evaluated on the basis of performance criteria such as sensitivity, specificity and predictive value.

The results of this research demonstrate the effectiveness of ensemble regression techniques in improving the accuracy of epidemiological forecasts, and offer new perspectives for public health intervention in affected regions.

Keywords : *Chikungunya, Ensemble regression, Random Forest, XGBoost, Machine learning, Climatic data, Epidemiological forecast*

Introduction Générale

Le chikungunya est une maladie virale transmise par la piqûre de moustiques infectés, principalement des espèces Aedes aegypti et Aedes albopictus.Le 2 août 2024, le Centre européen de contrôle et de prévention des maladies (ECDC) a indiqué qu'environ **350 000** cas de maladie à virus chikungunya (CHIKVD) et plus de **140 décès** ont été signalés dans le monde en 2024, Ces cas proviennent de 21 pays d'Amérique, d'Asie, d'Afrique et d'Europe [7].le chikungunya se manifeste par des symptômes tels que fièvre, douleurs articulaires, maux de tête, douleurs musculaires, gonflements articulaires, et éruptions cutanées. Bien que les décès dus au chikungunya soient rares, le virus peut provoquer de graves complications, en particulier chez les personnes âgées ou celles souffrant de maladies chroniques. Une détection précoce est cruciale pour prévenir la propagation de la maladie.

L'objectif de ce mémoire est de développer un modèle prédictif du chikungunya, en utilisant des approches de régression d'ensemble issues de l'intelligence artificielle. Compte tenu du manque de données suffisantes pour le Tchad, nous avons étendu notre étude aux données du Brésil et du Paraguay, où des cas de chikungunya ont également été signalés. Cette approche nous permet de tirer parti d'un ensemble de données plus vaste et diversifié pour améliorer la précision de nos prévisions.

Nous nous posons donc la question suivante : comment utiliser les techniques d'apprentissage automatique, en particulier l'approche ensembliste, pour prédire et analyser les épidémies de chikungunya ?

Les objectifs spécifiques de ce travail sont les suivants :

- Comprendre les concepts liés à la régression d'ensemble et au chikungunya ;
- Prédire les épidémies de chikungunya en se basant sur les données climatiques et les cas rapportés au Tchad, au Brésil, et au Paraguay ;
- Analyser la dynamique du chikungunya dans ces régions et proposer des stratégies d'intervention.

Le reste du document est structuré comme suit : le premier chapitre est dédié à l'épidémiologie du chikungunya avec un accent particulier sur les cas au Tchad, au Brésil et au Paraguay. Le deuxième chapitre introduit l'apprentissage automatique dans le contexte des maladies infectieuses. Le troisième chapitre présente la conception générale et détaillée du modèle proposé. Dans le quatrième chapitre, nous illustrons l'efficacité de notre modèle à travers des résultats de simulation. Enfin, le mémoire se conclut par une synthèse des résultats et des perspectives pour des recherches futures.

Chapitre 1

Épidémiologie de la Chikungunya

Le chikungunya est une maladie virale transmise à l'homme par des moustiques infectés par le virus du chikungunya. Les moustiques impliqués dans la transmission sont Aedes aegypti et Aedes albopictus [14].

1.1 Origine de la Chikungunya

La fièvre **Chikungunya** est une maladie virale décrite pour la première fois en 1952 lors d'une épidémie dans le sud de la **Tanzanie**. Le nom vient d'un mot de la langue *Makonde*, parlée dans le sud-est de la Tanzanie et le nord du Mozambique, qui signifie "*devenir contorsionné*" ou "*ce qui se plie*". Le virus a été isolé pour la première fois en Thaïlande en 1958.[18]

Évolution géographique et épidémiologique

En **avril 2005**, il a été confirmé que le CHIKV était à l'origine d'une épidémie de maladie ressemblant à la dengue sur les îles Comores, situées au large de la côte est du Mozambique septentrional ; il s'agissait de la première émergence connue du CHIKV dans la région du sud-ouest de l'océan Indien. En raison de similitudes cliniques, cette épidémie a d'abord été suspectée d'être causée par le virus de la dengue, soulignant le fait que la maladie CHIKV est souvent mal diagnostiquée et que le nombre réel de cas dans une région donnée peut être sous-estimé. Peu après, les premiers cas de CHIKV ont été signalés à Mayotte, à Maurice et sur l'île française de La Réunion. Le nombre de cas dans ces régions a rapidement augmenté, notamment en raison de taux d'attaque atteignant **35%** à **75%**. À la fin de l'année **2005**, après une période apparente d'environ 32 ans pendant laquelle le CHIKV n'a pas été détecté, l'Inde a signalé des cas de maladie à CHIKV dans de nombreux États, le nombre officiel de cas suspects atteignant finalement plus de 1,3 million. L'épidémie de CHIKV a continué à se propager, provoquant d'importantes flambées au Sri Lanka et dans de nombreux autres pays d'Asie du Sud-Est. Au cours de cette épidémie, le CHIKV a été introduit dans des pays où il n'est pas endémique par des voyageurs virémiques, et la transmission autochtone du CHIKV a été observée pour la première fois dans de nombreux pays, dont l'Italie, la France, la Nouvelle-Calédonie, la Papouasie-Nouvelle-Guinée, le Bhoutan et le Yémen. La propagation rapide et explosive du CHIKV a incité l'Organisation panaméricaine de la santé (OPS) et les Centers for Disease Control and Prevention (CDC) à publier un guide de préparation qui prévoyait de futures épidémies potentielles de CHIKV dans les Amériques. Cette prédiction s'est maintenant concrétisée,

puisqu'en décembre 2013, l'Organisation mondiale de la santé (OMS) a signalé la première transmission locale du CHIKV dans l'hémisphère occidental, sur l'île caribéenne de Saint-Martin. Le 18 juillet 2014, le CHIKV avait provoqué plus de 440 000 cas de maladie dans plus de 20 pays des Caraïbes, d'Amérique centrale et d'Amérique du Sud (Fig. 1). En outre, les CDC ont signalé plus de 230 cas importés d'infection à CHIKV sur le territoire continental des États-Unis, ainsi que des cas acquis localement en Floride. Ainsi, en moins de 10 ans, le CHIKV s'est propagé depuis les côtes du Kenya dans l'océan Indien, le Pacifique et les Caraïbes, provoquant des millions de cas de maladie dans plus de 50 pays. En d'autres termes, le CHIKV est redevenu un véritable agent pathogène mondial.

FIGURE 1.1 – CHIKV in the Western Hemisphere.

1.2 Agent Pathogène

Le virus du chikungunya (CHIKV), un alphavirus transmis par les moustiques, dont le génome est constitué d'un ARN monocaténaire à sens positif de ∼12 kb [13].

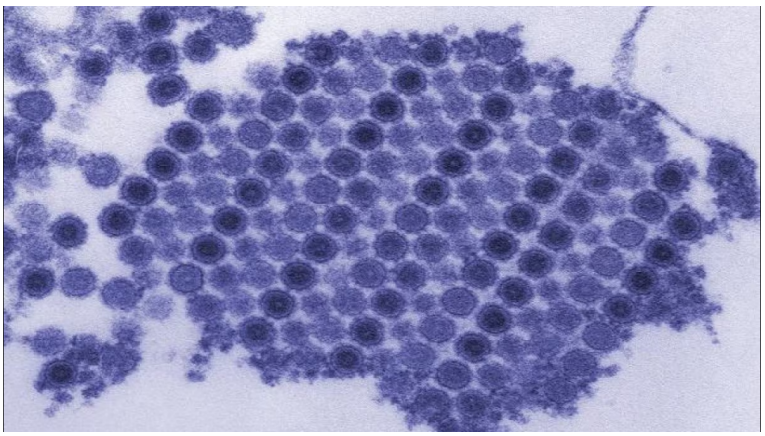

FIGURE 1.2 – Electron microscopic image of chikungunya virus

1.3 Mode de Transmission

Comprendre les mécanismes de transmission du virus chikungunya est essentiel pour développer des stratégies de prévention efficaces. Cette section examine les principaux vecteurs de la maladie, les conditions environnementales qui favorisent la propagation du virus et les dynamiques de transmission entre les hôtes humains et animaux.

1.3.1 Vecteurs : Les moustiques Aedes

Le virus du chikungunya est un alphavirus, similaire aux virus de Mayaro et de Ross River, appartenant à la famille *Togaviridae*, au genre *Alphavirus*. Le virus du chikungunya est principalement transmis à l'homme par la piqûre d'un moustique infecté, principalement *Aedes aegypti* (Voir Figure 1.3) et *Ae. albopictus* .

FIGURE 1.3 – Aedes aegypti mosquito full of blood

1.3.2 La transmission

Le **CHIKV** se transmet selon deux cycles différents :

- **Cycle urbain** : transmission de l'homme au moustique.
- **Cycle sylvatique** : transmission de l'animal au moustique, puis à l'homme [8].

Le cycle sylvatique est la principale forme de transmission en Afrique [8]. Ailleurs, dans les zones plus densément peuplées, le CHIKV se maintient principalement dans un cycle urbain, dans lequel les humains sont les principaux hôtes et les moustiques du genre *Aedes* les vecteurs [8] (voir figure1.4). bien que *Ae. aegypti* continue d'être un vecteur viral important, comme on l'a vu lors de la flambée épidémique dans les Caraïbes en 2013 [8]. La transmission verticale de la mère à l'enfant a été postulée pour expliquer les incidences postérieures à 2005 [8], étant particulièrement délétère lorsque :

- la mère est infectée jusqu'à quatre jours après l'accouchement [8],

bien que cette hypothèse ait été contestée [8].

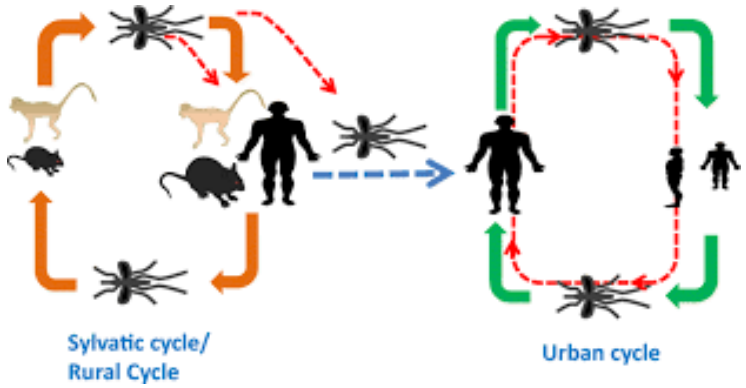

FIGURE 1.4 – Transmission mode Chikungunya

1.4 Symptômes et Diagnostic

La reconnaissance des symptômes et l'établissement d'un diagnostic précis sont essentiels pour la gestion efficace des cas de chikungunya. Cette section examine les manifestations cliniques typiques de l'infection par le virus chikungunya ainsi que les méthodes diagnostiques utilisées pour identifier la maladie.

1.4.1 Symptômes

Chez les patients symptomatiques, la maladie à **CHIKV** se déclare généralement 4 à 8 jours (entre 2 et 12 jours) après la piqûre d'un moustique infecté. Elle se caractérise par une brusque poussée de fièvre, souvent accompagnée de fortes douleurs articulaires. Les douleurs articulaires sont souvent invalidantes et durent généralement quelques jours, mais peuvent être prolongées et durer des semaines, des mois, voire des années. D'autres signes et symptômes courants sont le gonflement des articulations, les douleurs musculaires, les maux de tête, les nausées, la fatigue et les éruptions cutanées. Comme ces symptômes se confondent avec ceux d'autres infections, notamment celles dues aux virus de la dengue et du Zika, les cas peuvent être mal diagnostiqués. En l'absence de douleurs articulaires importantes, les symptômes des personnes infectées sont généralement légers et l'infection peut passer inaperçue.

La plupart des patients se rétablissent complètement de l'infection ; toutefois, des cas occasionnels de complications oculaires, cardiaques et neurologiques ont été signalés dans le cadre d'infections par le **CHIKV**. Les patients situés aux extrémités du spectre d'âge sont plus exposés à une maladie grave. Les nouveau-nés infectés pendant l'accouchement et les personnes âgées souffrant de pathologies sous-jacentes peuvent devenir gravement malades et l'infection par le **CHIKV** peut augmenter le risque de décès [19].

Une fois qu'une personne est guérie, les données disponibles suggèrent qu'elle est probablement immunisée contre les infections futures [2].

1.4.2 Méthodes de diagnostic

Le diagnostic peut être retardé en raison de la confusion possible des symptômes avec ceux de la dengue ou du Zika. Les tests ***immuno-enzymatiques*** (ELISA) peuvent être utilisés pour confirmer la présence d'anticorps **anti-CHIKV**, les niveaux d'anticorps IgM étant les plus élevés trois à cinq semaines après l'infection et persistant jusqu'à deux mois. La PCR peut également être utilisée pour génotyper le virus [13].

1.5 Méthodes de Contrôle et Traitement

La gestion efficace de l'épidémie de chikungunya repose sur une combinaison de stratégies de contrôle des vecteurs et d'interventions médicales. Cette section explore les diverses approches utilisées pour prévenir la transmission du virus et traiter les symptômes chez les patients infectés.

1.5.1 Méthodes de contrôle

La prévention de l'infection en évitant les piqûres de moustiques est la **meilleure protection**. Les patients suspectés d'être infectés par le **CHIKV** doivent éviter les piqûres de moustiques pendant la première semaine de la maladie afin d'empêcher la transmission aux moustiques, qui peuvent à leur tour infecter d'autres personnes.

La principale méthode pour réduire la transmission du **CHIKV** consiste à contrôler les moustiques vecteurs. Pour ce faire, il faut mobiliser les communautés, qui jouent un rôle essentiel dans la réduction des sites de reproduction des moustiques en :

- vidant et en nettoyant chaque semaine les récipients contenant de l'eau,
- éliminant les déchets,
- soutenant les programmes locaux de lutte contre les moustiques.

Pendant les épidémies, des insecticides peuvent être :

- pulvérisés pour tuer les moustiques adultes volants,
- appliqués sur les surfaces à l'intérieur et autour des conteneurs où les moustiques se posent,
- utilisés pour traiter l'eau dans les conteneurs afin de tuer les larves immatures.

Les autorités sanitaires peuvent également prendre des mesures d'urgence pour contrôler la population de moustiques.

Pour se protéger pendant les épidémies de **chikungunya**, il est conseillé de :

- porter des vêtements qui minimisent l'exposition de la peau aux vecteurs qui piquent pendant la journée,
- utiliser des moustiquaires aux fenêtres et aux portes pour empêcher les moustiques de pénétrer dans les maisons,
- appliquer des répulsifs sur la peau exposée ou sur les vêtements en respectant scrupuleusement les instructions figurant sur l'étiquette du produit.

Les répulsifs doivent contenir du **DEET**, de l'**IR3535** ou de l'**icaridine** [**?**].

1.5.2 Options de traitement

Le traitement actuel vise à atténuer la gravité des symptômes plutôt qu'à guérir la maladie. Le traitement repose principalement sur l'utilisation d'**antipyrétiques** et d'**AINS**. Cependant, aucune étude n'a évalué systématiquement l'efficacité de ces traitements, et les symptômes peuvent disparaître sans intervention. L'utilisation de **corticostéroïdes** pour le traitement de la phase aiguë a connu un succès mitigé et est utilisée avec hésitation en raison de la possibilité d'aggravation des symptômes après le traitement. Il est particulièrement important de maintenir des niveaux de liquide adéquats.

Il existe également des preuves émergentes que les médicaments qui entravent le transport du cholestérol, tels que les composés **amphiphiles cationiques de classe II U18666A** et l'**imipramine**, peuvent être efficaces contre la fusion membranaire du **CHIKV**, et ont un potentiel d'action contre d'autres arbovirus.

Pour les arthralgies chroniques graves, des **antirhumatismaux modificateurs de la maladie (ARMM)**, notamment le **méthotrexate**, l'**hydroxychloroquine** ou la **sulfasalazine**, ont été proposés. Comme pour les traitements aigus, l'efficacité systématique des **DMARD** pour le traitement chronique est inconnue, bien que des rapports décrivent des résultats positifs avec une cessation des symptômes dans les 4 à 6 mois [8].

1.6 Cas du Tchad

L'analyse spécifique des cas de chikungunya au Tchad permet de comprendre l'impact de cette maladie dans un contexte régional spécifique. Cette section examine les Points saillant, la mise en contexte du problème dans ce pays d'Afrique central.

1.6.1 Point saillants

le 3 Septembre 2020, 927 cas ont été notifiés, tous pris en charge en ambulatoire, sans aucun décès. Le 3 septembre 2020, le cumul atteignait 13 488 cas, toujours sans décès, avec de nouveaux cas suspects signalés à **Biltine** (10) et **Adré** (2).Le 2 octobre 2020, **415 cas** ont été notifiés, répartis comme suit : 247 à **Abéché**, 165 à **Biltine**, 3 à **Gozbeida**, et 0 à **Abdi**, sans aucun décès. à cette date , le cumul atteignait **34 052 cas**, avec **un** décès. Tous les patients ont été pris en charge en ambulatoire [6].

FIGURE 1.5 – Region infecté au Tchad

1.6.2 Contexte

En **juillet 2020**, au Tchad, des cas d'une maladie localement appelée Kourgnalé, présentant des symptômes tels que fièvre, céphalées et douleurs articulaires, ont été signalés à Abéché. En août, l'augmentation des cas a conduit à des analyses confirmant la présence du virus Chikungunya le **12 août 2020**. Entre le 14 août et le 3 septembre dans le rapport de L'OMS, **13 488** cas ont été enregistrés sans aucun décès [6]. Les femmes et les personnes âgées de 15 ans et plus étaient les plus touchées. Plus de **75 %** des patients présentaient des symptômes graves, et un tiers souffrait d'éruptions cutanées.

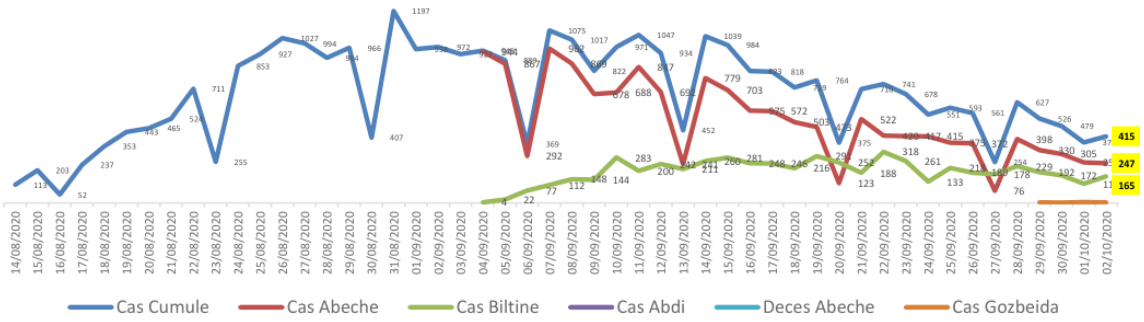

FIGURE 1.6 – Evolution journalière des cas et décès du Chikungunya

1.7 Cas du Brésil

Le Brésil est le pays le plus grand et le plus peuplé d'Amérique latine, avec une importante population sensible au CHIKV, ainsi qu'un climat approprié et d'abondantes populations de vecteurs *Ae. aegypti*. Le CHIKV circule localement au Brésil depuis **2014**, les premiers cas étant principalement limités au **Nord-Est**.Depuis 2016, le Brésil est l'épicentre des épidémies de chikungunya dans les Amériques avec **1 659 167** cas, le nombre le plus élevé rapporté dans la région. Contrairement à d'autres pays et territoires des Amériques, le Brésil connaît des épidémies annuelles de chikungunya [5]. cette section examine les Points saillant, la mise en contexte du problème dans ce pays.

1.7.1 Point saillants

La figure 1.7 montre les cas cumulés de chikungunya (c'est-à-dire les cas suspectés et confirmés en laboratoire). Dans les 26 États brésiliens et le district fédéral déclarés au ministère brésilien de la santé entre mars 2013 et juin 2023. (b) Distribution spatio-temporelle des lignées du virus du chikungunya dans les pays et territoires des 26 États brésiliens et du district fédéral. La lignée du chikungunya en circulation a été déterminée sur la base des années pour lesquelles au moins un génome a été séquencé et déposé dans GenBank jusqu'au 17 août 2023. AC = Acre. AL = Alagoas. AM = Amazonas. AP = Amapá. BA = Bahia. CE = Ceará. ES = Espírito Santo. DF = Distrito Federal (district fédéral). GO = Goiás. MA = Maranhão. MG = Minas Gerais. MS = Mato Grosso do Sul. MT = Mato Grosso. PA = Pará. PB = Paraíba. PE = Pernambouc. PI = Piauí. PR = Paraná. RJ = Rio de Janeiro. RN = Rio Grande do Norte. RO = Rondônia. RR = Roraima. RS = Rio Grande do Sul. SC = Santa Catarina. SE = Sergipe. SP = São Paulo. TO = Tocantins. Km = kilomètres. ECSA-American, sous-lignée est-centrale-sud-africaine-américaine [5].

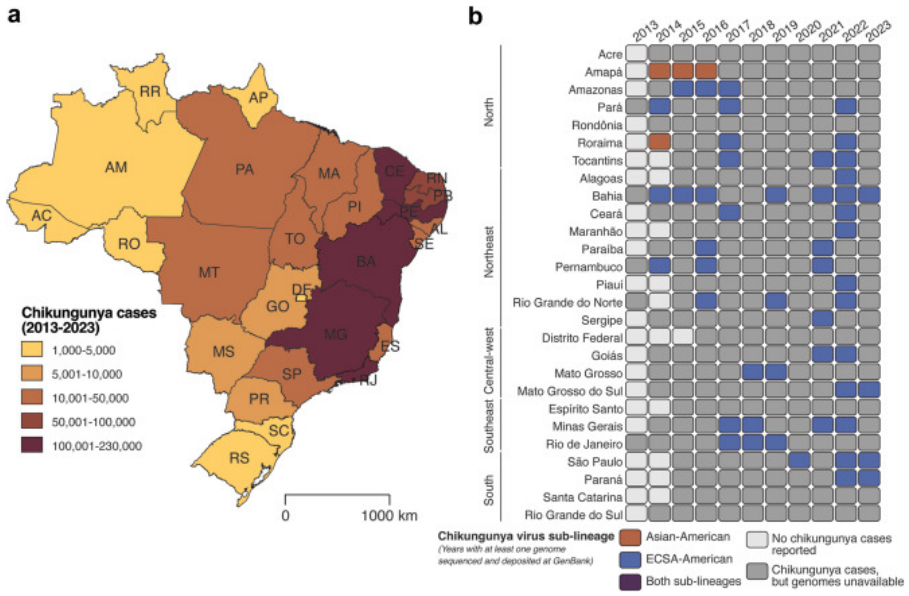

FIGURE 1.7

1.7.2 Contexte

Le chikungunya a ressurgi en 2022-23 après une période d'accalmie, touchant gravement le **Paraguay** avec une incidence cumulative de **1216** cas pour 100 000 habitants en 2023, et le Brésil, notamment dans l'État

de Minas Gerais avec **395** cas pour 100 000 habitants. Depuis son introduction au Brésil en 2014, avec **3,6 millions** de cas signalés à la PAHO/WHO, la maladie s'est déplacée du Nord-Est vers le Sud-Est, où en 2023, **30 724 cas** ont été rapportés en seulement 10 semaines, soit deux fois plus qu'en 2022. Le taux de reproduction du virus a atteint des valeurs élevées, entre 1,5 et 2,5, avec des pics en 2018 et 2022 [1].

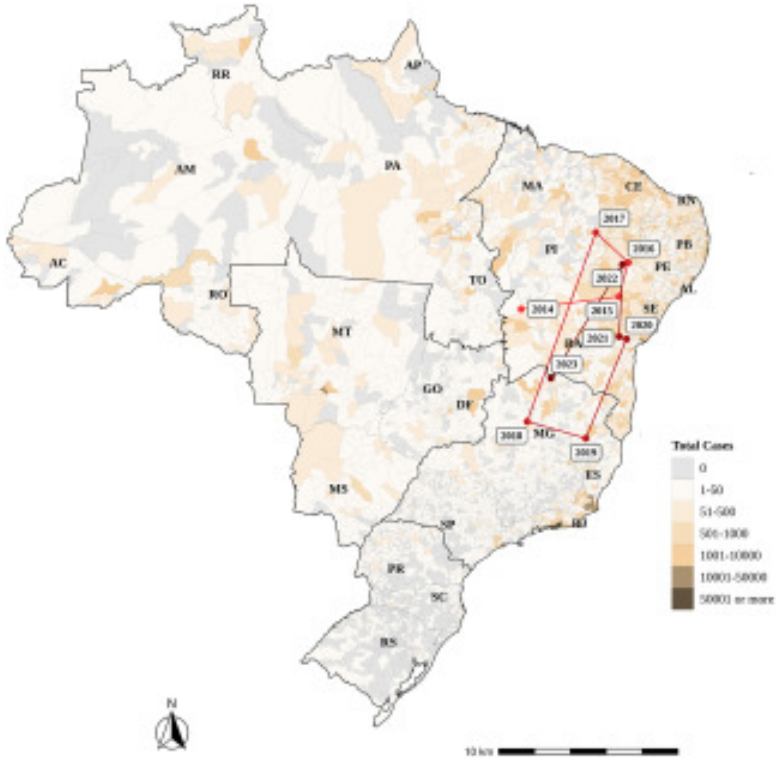

FIGURE 1.8

1.8 Cas du Paraguay

Des infections autochtones ont été détectées au **Paraguay** en 2015, et le CHIKV a été détecté dans le pays chaque année depuis cette date. Sur la base des infections suspectes à CHIKV signalées, le Paraguay a connu quatre vagues épidémiques, en **2015**, **2016**, **2018** et **2023**, toutes associées aux mois d'été. Du 2 octobre 2022 au 10 avril 2023, un total de **118 179** infections suspectes et confirmées ont été signalées, dont **3 510** cas-patients hospitalisés et 46 décès. Les nouveau-nés ont représenté 0,3 % (n = 162) de ces cas et 8 décès. En outre, **294** cas suspects de méningo-encéphalite aiguë ont été signalés, dont 125 (43%) ont été attribués au CHIKV [9].

1.8.1 Point saillants

Bien que les températures minimales annuelles soient restées stables au Paraguay au cours des 40 dernières années, les températures moyennes et maximales annuelles ont augmenté régulièrement, et la résurgence rapide et importante du CHIKV en 2022 a coïncidé avec les températures moyennes les plus élevées signalées. Avant 2022, les infections confirmées étaient limitées aux districts de *Central*, *Paraguarí* et *Amambay* ; le district de Central dominait les rapports. Après la résurgence virale de 2022, des infections confirmées ont été signalées dans tous les districts.

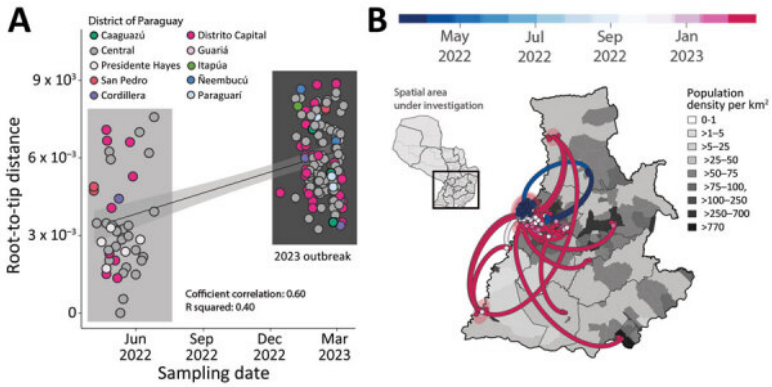

FIGURE 1.9 – A) L'analyse examine les distances génétiques entre les échantillons de virus et leurs dates d'échantillonnage, avec des intervalles de confiance à 90 %. Les couleurs montrent les lieux d'origine des échantillons. B) La propagation du virus CHIKV ECSA au Paraguay est cartographiée, les cercles indiquant les points clés de la phylogénie, montrant comment le virus s'est diffusé géographiquement au fil du temps.

1.8.2 Contexte

Cas de chikungunya rapportés chaque semaine (zone grise), incidence normalisée pour **100 000** personnes (ligne bleue) et **décès** cumulés voir Figure 1.10.

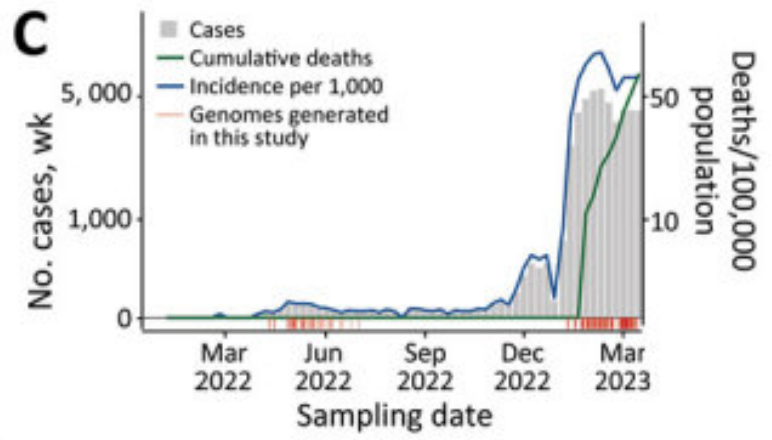

FIGURE 1.10 – Weekly reported chikungunya cases

Chapitre 2

Revue de la Littérature et Concepts de Base

Dès **1997**, le potentiel du **data mining** pour améliorer les problèmes dans le domaine médical avait été identifié par l'Organisation mondiale de la santé (**OMS**) (Gulbinat, 1997). L'utilité de la détection de connaissances à partir des dépôts de données médicales a été soulignée par l'OMS, car elle bénéficie au diagnostic médical et à la prédiction. Le data mining est un processus de découverte de connaissances utiles à partir de bases de données pour construire une structure (c'est-à-dire un modèle ou un schéma) qui peut interpréter de manière significative les données. Le data mining est le processus de découverte de schémas et de connaissances intéressants à partir d'une grande quantité de données (Han et al., 2001). Le data mining utilise de nombreuses techniques d'apprentissage automatique pour découvrir des schémas cachés dans les données. Ces techniques peuvent être réparties en trois catégories principales : les techniques d'**apprentissage supervisé**, les techniques d'**apprentissage non supervisé** et les techniques d'**apprentissage semi-supervisé** (Huang al., 2014). Voir figure 2.1. Les systèmes experts développés par des techniques d'apprentissage automatique peuvent être utilisés pour aider les médecins dans le **diagnostic** et la **prédiction des maladies** (Kononenko, 2001). En raison de l'importance du diagnostic des maladies pour l'humanité, plusieurs études ont été menées sur le développement de méthodes pour leur classification [11].

2.1 Apprentissage Automatique

Depuis leur évolution, les humains ont utilisé de nombreux types d'outils pour accomplir diverses tâches de manière plus simple. La créativité du cerveau humain a conduit à l'invention de différentes machines. Ces machines ont facilité la vie humaine en permettant aux gens de répondre à divers besoins de la vie, y compris le voyage, les industries et l'informatique. Et l'apprentissage automatique en fait partie. Selon Arthur Samuel, l'**apprentissage automatique** est défini comme le domaine d'étude qui donne aux ordinateurs la capacité d'apprendre sans être explicitement programmés. *Arthur Samuel* était célèbre pour son programme de jeu de dames. L'apprentissage automatique (**ML**) est utilisé pour apprendre aux machines comment gérer les données de manière plus efficace. Parfois, après avoir examiné les données, nous ne pouvons pas interpréter les informations extraites des données. Dans ce cas, nous appliquons l'apprentissage automatique. Avec l'abondance des ensembles de données disponibles, la demande pour l'apprentissage automatique est en hausse. De nom-

breuses industries appliquent l'apprentissage automatique pour extraire des données pertinentes. Le but de l'apprentissage automatique est d'*apprendre à partir des données* [10].

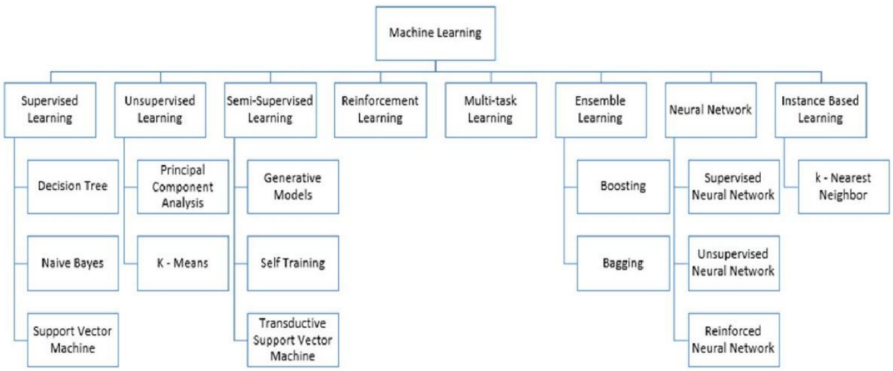

FIGURE 2.1 – Machine Learning Tree

2.1.1 Techniques d'Apprentissage Automatique

Voici un aperçu rapide de certains des algorithmes couramment utilisés en apprentissage automatique (ML)

2.1.1.1 Apprentissage Supervisé

Les algorithmes d'**apprentissage supervisé** sont ceux qui nécessitent une assistance externe. L'ensemble de données d'entrée est divisé en ensemble d'entraînement et ensemble de test. L'ensemble d'entraînement a une variable de sortie qui doit être prédite ou classée. Tous les algorithmes apprennent un certain type de schémas à partir de l'ensemble d'entraînement et les appliquent à l'ensemble de test pour la prédiction ou la classification. Le flux de travail des algorithmes d'apprentissage supervisé est donné dans la figure 2.2 ci-dessous [10].

Exemple d'apprentissage supervisé :

- **Arbre de Décision** : est un graphe pour représenter les choix et leurs résultats sous forme d'arbre.
- **Naïve Bayes** : C'est une technique de classification basée sur le théorème de Bayes avec une hypothèse d'indépendance entre les prédicteurs. En termes simples, un classificateur Naïve Bayes suppose que la présence d'une caractéristique particulière dans une classe est indépendante de la présence de toute autre caractéristique.

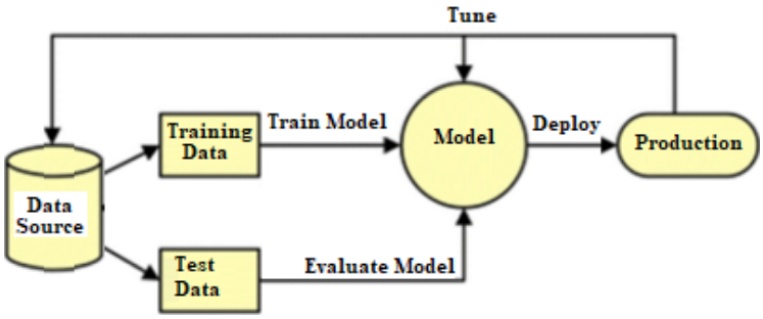

FIGURE 2.2 – Flux de travail de l'apprentissage supervisé

2.1.1.2 Apprentissage Non Supervisé

Contrairement à l'apprentissage supervisé ci-dessus, il n'y a pas de réponses correctes et il n'y a pas de professeur. Les algorithmes sont laissés à leurs propres dispositifs pour découvrir et présenter la structure intéressante dans les données. Les algorithmes d'apprentissage non supervisé apprennent quelques caractéristiques à partir des données. Lorsque de nouvelles données sont introduites, elles utilisent les caractéristiques apprises précédemment pour reconnaître la classe des données. Il est principalement utilisé pour le *clustering* et la réduction des *features* [10].

Exemple d'apprentissage non supervisé :

– **Clustering K-means** : est l'un des algorithmes d'apprentissage non supervisé les plus simples qui résout le problème bien connu du clustering. La procédure suit une manière simple et facile de classer un ensemble de données donné par un certain nombre de clusters.

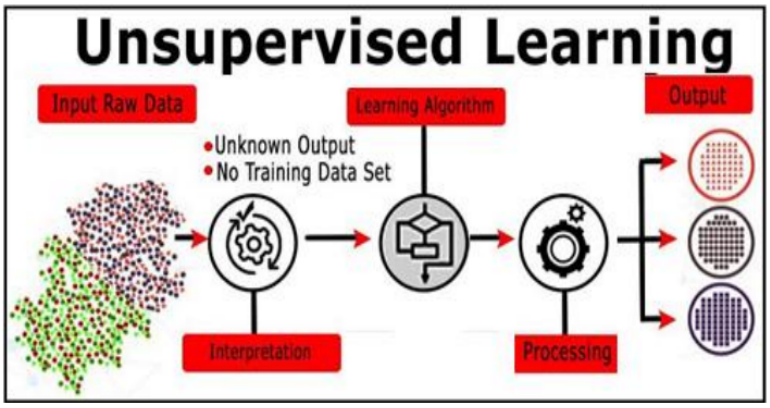

FIGURE 2.3 – Apprentissage Non Supervisé

2.1.1.3 Apprentissage Semi-Supervisé

L'**apprentissage semi-supervisé** est une combinaison des méthodes d'apprentissage *supervisé* et *non supervisé*. Il peut être fructueux dans ces domaines de l'apprentissage automatique et du data mining où les données non étiquetées sont déjà présentes et obtenir les données étiquetées est un processus fastidieux [10].

2.2 Introduction au Deep Learning

Le *Deep Learning* est une branche de l'intelligence artificielle (IA) qui s'appuie sur des réseaux de neurones artificiels pour modéliser des relations complexes dans les données. Ce domaine a connu une croissance exponentielle grâce aux avancées technologiques en matière de calcul et aux grandes quantités de données disponibles. Ce cours explore les concepts fondamentaux du Deep Learning.

2.2.1 Les bases du Deep Learning

2.2.1.1 Réseaux de neurones

Les *neurones artificiels*, inspirés du fonctionnement des neurones biologiques, sont les unités de base des réseaux de neurones. Chaque neurone reçoit des entrées, les traite à l'aide d'une fonction d'activation, puis transmet une sortie.

Parmi les caractéristiques nous avons :

- **Couches et profondeur** : Les neurones sont organisés en couches (entrée, cachées, sortie). Un réseau est dit "profond" lorsqu'il possède plusieurs couches cachées, permettant ainsi de capturer des abstractions de plus en plus complexes des données d'entrée.
- **Poids et apprentissage** : Chaque connexion entre deux neurones est associée à un poids, qui détermine l'influence de l'entrée sur la sortie. Ces poids sont ajustés lors de la phase d'apprentissage pour minimiser l'erreur du modèle.

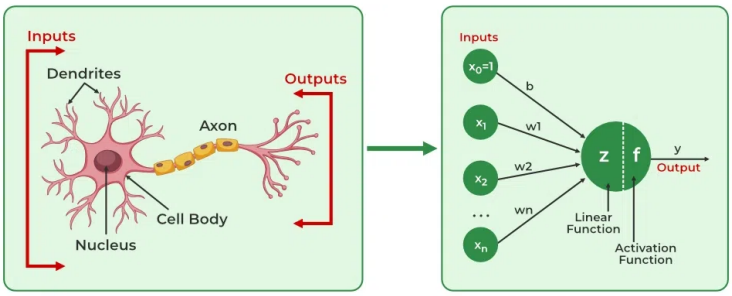

FIGURE 2.4 – Neurone Biologique et Artificiel

2.2.1.2 Fonctionnement du réseau

Le Deep Learning repose souvent sur l'apprentissage supervisé, où le modèle est entraîné à partir d'un ensemble de données étiquetées.

- **Propagation avant** : Les données d'entrée sont transmises couche par couche, chaque neurone calculant une activation à partir de ses entrées pondérées.
- **Rétropropagation** : L'algorithme de rétropropagation ajuste les poids en fonction de l'erreur commise par le modèle, en propagant cette erreur en sens inverse, de la sortie vers l'entrée.

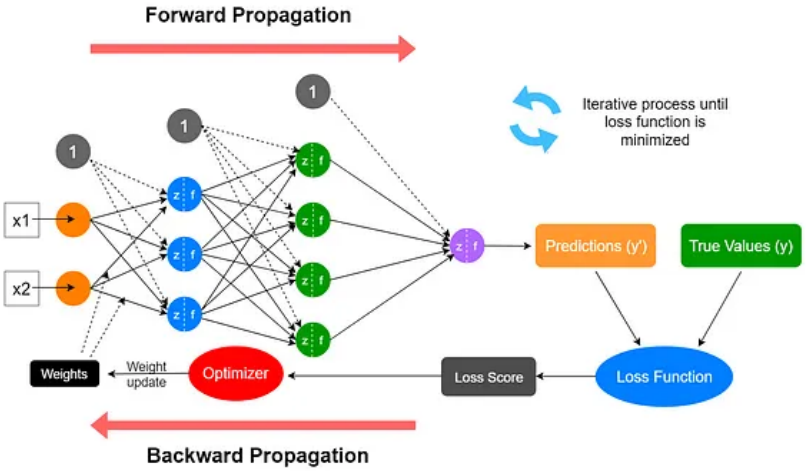

FIGURE 2.5 – Fonctionnement réseau de neurone

2.2.2 Fonctions d'activation

Les fonctions d'activation introduisent de la non-linéarité dans les réseaux de neurones, permettant de modéliser des relations complexes.

Exemples de fonctions d'activation

- **ReLU (Rectified Linear Unit)** :

$$f(x) = \max(0, x)$$

  ReLU est couramment utilisée car elle permet un apprentissage efficace tout en réduisant les risques de *vanishing gradient*[1].

- **Sigmoid** :

$$f(x) = \frac{1}{1 + e^{-x}}$$

  La fonction Sigmoid est utilisée pour normaliser la sortie entre 0 et 1, particulièrement utile pour les tâches de classification binaire.

- **Tanh (Tangente hyperbolique)** :

$$f(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$

  Tanh normalise la sortie entre -1 et 1, ce qui peut conduire à une convergence plus rapide que la Sigmoid dans certains cas.

2.2.3 Entraînement des modèles

2.2.3.1 Fonction de coût

- **Évaluation de l'erreur** : La fonction de coût mesure l'écart entre les prédictions du modèle et les vraies étiquettes, guidant ainsi l'ajustement des poids.
- **Exemples de fonctions de coût** :
  - **Erreur quadratique moyenne (MSE)** :

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

    Utilisée pour les problèmes de régression.

  - **Erreur quadratique moyenne racine (RMSE)** :

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

    Le RMSE est une mesure courante de l'erreur moyenne entre les valeurs prédites et les valeurs réelles. Il est particulièrement utile pour comparer les performances de différents modèles, car il conserve les mêmes unités que les données. **Valeurs possibles** :

1. Le *vanishing gradient* est un problème courant dans les réseaux de neurones profonds, où les gradients deviennent extrêmement petits, empêchant les poids des couches précédentes de se mettre à jour correctement, ce qui ralentit considérablement l'apprentissage.

- **RMSE = 0** : Indique une prédiction parfaite, où les valeurs prédites sont exactement égales aux valeurs réelles.
- **RMSE > 0** : Plus le RMSE est élevé, plus l'erreur moyenne est grande. Un RMSE élevé suggère que le modèle a du mal à prédire les valeurs réelles avec précision.

on cherche à minimiser le RMSE pour améliorer la précision du modèle.

- **R-carré (R^2)** :

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Le R^2 score, également appelé coefficient de détermination, mesure la proportion de la variance des données qui est expliquée par le modèle. Un score proche de 1 indique que le modèle explique bien les données, tandis qu'un score proche de 0 indique le contraire.

2.2.3.2 Optimisation

L'optimisation est une étape cruciale dans l'entraînement des réseaux de neurones, où l'objectif est de minimiser la fonction de coût en ajustant les poids du réseau. Cela se fait en utilisant des algorithmes d'optimisation qui guident le processus d'ajustement des poids pour réduire l'erreur entre les prédictions du modèle et les valeurs réelles.

Descente de gradient : L'algorithme le plus couramment utilisé pour l'optimisation en Deep Learning est la descente de gradient. L'idée principale est de mettre à jour les poids du réseau dans la direction opposée au gradient de la fonction de coût par rapport aux poids. Ce processus est itératif et continue jusqu'à ce que la fonction de coût atteigne un minimum local ou global.

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta J(\theta_t)$$

où :

- θ_t représente les poids du réseau à l'itération t,
- η est le taux d'apprentissage (un hyperparamètre qui contrôle la taille des mises à jour des poids),
- $\nabla_\theta J(\theta_t)$ est le gradient de la fonction de coût $J(\theta)$ par rapport aux poids θ.

2.2.4 Applications du Deep Learning

Vision par ordinateur

- **Tâches principales** : Classification d'images, détection d'objets, segmentation sémantique.
- **Exemple d'architecture** : Les réseaux de neurones convolutifs (CNN) sont la norme pour les tâches de vision par ordinateur, offrant d'excellentes performances sur des tâches complexes comme la reconnaissance d'images.

Traitement du langage naturel (NLP)

- **Tâches principales** : Traduction automatique, résumé de texte, chatbots, analyse de sentiments.
- **Exemple d'architecture** : Les réseaux récurrents (RNN) et les transformers (comme BERT) sont largement utilisés pour modéliser des séquences textuelles et capturer des dépendances à long terme.

2.3 Ensemble Learning ou Apprentissage par Ensemble

Les *systèmes à classificateurs multiples*, ou *systèmes d'ensemble*, ont gagné en popularité au sein de la communauté de l'intelligence artificielle et de l'apprentissage automatique au cours des deux dernières décennies. Initialement développés pour réduire la variance et améliorer la précision des décisions automatisées, ces systèmes se sont avérés efficaces et polyvalents dans divers domaines. Ils sont utilisés pour résoudre des problèmes tels que la sélection de caractéristiques, l'estimation de la confiance, et la gestion de données déséquilibrées. Bien que la recherche sur les systèmes d'ensemble soit relativement récente, la prise de décision collective, semblable aux systèmes d'ensemble, fait partie de notre quotidien depuis longtemps, comme en témoignent la démocratie et le système judiciaire [15].

2.3.1 Définition

L'**apprentissage par ensemble** est le processus par lequel plusieurs **modèles**, sont générés et combinés stratégiquement pour résoudre un problème particulier d'intelligence computationnelle. L'apprentissage par ensemble est principalement utilisé pour améliorer les performances d'un modèle ou réduire la probabilité de sélectionner un modèle médiocre [10].

2.3.2 principes de fonctionnement

Les **systèmes d'ensemble** sont inspirés par notre pratique quotidienne de consulter divers *experts* avant de prendre des décisions importantes, comme consulter plusieurs médecins avant une opération ou lire des avis avant un achat. Cette approche vise à augmenter notre confiance dans la décision finale en combinant différentes opinions.

En apprentissage automatique, les systèmes d'ensemble améliorent la **précision** en abordant les deux composantes principales de l'erreur de classification : le **biais** et la **variance**. Le biais reflète la précision d'un classificateur, tandis que la variance mesure la précision lors de l'utilisation de différents ensembles de données d'apprentissage. Généralement, un faible biais est associé à une variance élevée, et vice versa. Les systèmes d'ensemble cherchent à créer plusieurs classificateurs avec un biais similaire et à combiner leurs résultats, par exemple par **moyenne**, pour **réduire la variance** [15].

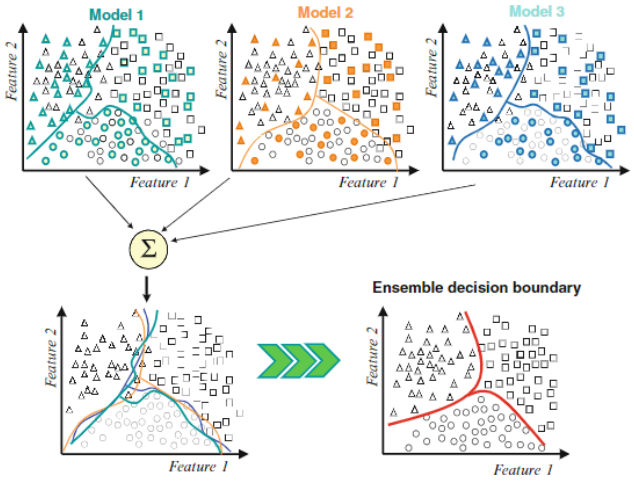

FIGURE 2.6 – Réduction de la variabilité à l'aide de systèmes d'ensemble

2.3.3 Techniques d'Ensemble Learning

Les premiers travaux sur les **systèmes d'ensemble** sont attribués à Dasarathy et Sheela (1979), qui ont exploré le **partitionnement de l'espace des caractéristiques** avec plusieurs **classificateurs** [*dasarathy1979*]. Dix ans plus tard, Hansen et Salamon ont démontré que des ensembles de **réseaux neuronaux** pouvaient améliorer les performances de **classification** [*hansen1990*]. Cependant, c'est Schapire qui a introduit le concept de **boosting**, montrant qu'un **classificateur puissant** pouvait être construit à partir de classificateurs ayant une erreur meilleure que celle d'une supposition aléatoire [*schapire1990*]. La théorie du boosting a conduit à l'algorithme **AdaBoost**, largement utilisé pour des problèmes de **classes multiples** et de **régression** [*freund1997*].

Depuis ces travaux initiaux, la recherche sur les systèmes d'ensemble a explosé, avec divers algorithmes apparaissant sous différents noms : **bagging** [*breiman1996*], *forêts aléatoires*, **mélanges d'experts** (MoE) [*jacobs1991*], **généralisation par empilement** [*wolpert1992*], et d'autres. Ces algorithmes diffèrent par la **sélection des données d'apprentissage**, la **procédure de génération des membres de l'ensemble**, et la **règle de combinaison** pour la décision finale. Ces trois éléments sont les **piliers** des systèmes d'ensemble.

Les systèmes d'ensemble sont principalement utilisés dans deux contextes : **sélection de classificateurs** et **fusion de classificateurs** [*kuncheva2004*]. Dans la sélection, chaque classificateur est un expert local, et celui formé avec les données les plus proches est choisi pour la décision finale. En revanche, dans la fusion, tous les classificateurs sont combinés pour obtenir un classificateur composite avec une **variance** plus faible, réduisant ainsi l'erreur [*breiman1996, freund1997*].

Construction d'un système d'ensemble

Trois stratégies doivent être choisies pour construire un système d'ensemble efficace. Nous les avons précédemment désignées comme les trois piliers des systèmes d'ensemble : (1) **échantillonnage/sélection des données** ; (2) **formation des classificateurs membres** ; et (3) **combinaison des classificateurs**.

2.3.3.1 Échantillonnage et sélection des données : Diversité

Dans les **systèmes d'ensemble**, la diversité des erreurs entre les membres est cruciale. Si tous les membres fournissent les mêmes résultats, leur combinaison n'apporte aucun bénéfice. Il est donc essentiel que les membres commettent des erreurs différentes, idéalement avec des résultats indépendants ou négativement corrélés.

La diversité peut être obtenue par diverses stratégies, comme l'utilisation de différents sous-ensembles des données d'apprentissage. Par exemple, le **bagging** utilise des répliques bootstrapées des données, tandis que le **boosting** se concentre sur les échantillons mal classés. D'autres méthodes incluent l'utilisation de sous-ensembles des **caractéristiques** disponibles ou différents **classificateurs** de base. Malgré l'importance reconnue de la diversité pour améliorer la performance des ensembles, aucune relation explicite entre la diversité et la précision de l'ensemble n'a été clairement établie [15].

2.3.3.2 Formation des classificateurs de membres

Au cœur de tout système basé sur un ensemble se trouve la stratégie utilisée pour former les membres individuels de l'ensemble. De nombreux algorithmes concurrents ont été développés pour former les classificateurs d'ensemble ; cependant, le *bagging* (et les algorithmes apparentés arc-x4 et Random Forest), le *boosting* (et ses

nombreuses variations), la *stack generalization* et le *hierarchical MoE* restent les approches les plus couramment employées.

2.3.3.3 Combinaison des membres de l'ensemble

La dernière étape des **systèmes d'ensemble** est la **combinaison des classificateurs** individuels. La stratégie de combinaison dépend du type de classificateurs utilisés. Par exemple, certains classificateurs comme les **Support Vector Machines (SVM)** produisent des sorties discrètes, pour lesquelles les règles de combinaison les plus courantes sont le **majority voting** (simple ou pondéré) et le **Borda count**. D'autres classificateurs, tels que le **multilayer perceptron** ou le **naïve Bayes classifier**, fournissent des sorties continues indiquant le soutien pour chaque classe. Pour ces classificateurs, une gamme plus large d'options est disponible, telles que les **arithmetic combinators** (somme, produit, moyenne) ou des **decision models** plus sophistiqués. Certains de ces combinateurs peuvent être appliqués immédiatement après l'apprentissage, tandis que des méthodes plus complexes comme **stacking** ou **hierarchical MoE** peuvent nécessiter une étape d'apprentissage supplémentaire. Nous examinerons brièvement ces approches.

Combining Class Labels (Combinaison des étiquettes de classe)

Cette sous-section se concentre sur les scénarios où les classifieurs produisent des étiquettes de classe discrètes.

- **Majority Voting (Vote majoritaire)** : La classe avec le plus de votes gagne. L'ensemble choisit la classe ω_c, si :
$$\sum_{t=1}^{T} d_{t,c} = \max_c \sum_{t=1}^{T} d_{t,c}$$
où $d_{t,c}$ est 1 si le classifieur t choisit la classe c, et 0 sinon.
- **Weighted Majority Voting (Vote majoritaire pondéré)** : Similaire au vote majoritaire, mais le vote de chaque classifieur est pondéré en fonction de ses performances estimées. L'ensemble choisit la classe c, si :
$$\sum_{t=1}^{T} w_t d_{t,c} = \max_c \sum_{t=1}^{T} w_t d_{t,c}$$
où w_t est le poids assigné au classifieur t.
- **Borda Count (Comptage Borda)** : Les classes sont classées par chaque classifieur, et la classe avec le score cumulé le plus élevé sur tous les classifieurs est sélectionnée.

Combining Continuous Outputs (Combinaison des sorties continues)

Cette sous-section traite des cas où les classifieurs produisent des sorties continues, souvent interprétées comme un support ou des probabilités pour chaque classe.

- **Algebraic Combiners (Combinateurs algébriques)** : Les supports pour chaque classe de différents classifieurs sont combinés à l'aide de diverses fonctions mathématiques. Le support total pour la classe ω_c est représenté comme suit :
$$\mu_c(x) = F[d_{1,c}(x), \ldots, d_{T,C}(x)]$$
où $F[]$ peut être la moyenne, la moyenne pondérée, le minimum, le maximum, la médiane, le produit ou la moyenne généralisée.

Les méthodes suivantes et leurs formules correspondantes sont :

- **Mean Rule (Règle de la moyenne)** : Le support pour une classe est la moyenne des supports de tous les classifieurs.

$$\mu_c(x) = \frac{1}{T}\sum_{t=1}^{T} d_{t,c}(x)$$

- **Weighted Average (Moyenne pondérée)** : Similaire à la règle de la moyenne, mais le support de chaque classifieur est pondéré.

$$\mu_c(x) = \frac{1}{T}\sum_{t=1}^{T} w_t d_{t,c}(x)$$

  où w_t est le poids du $t^{\text{ème}}$ classifieur.
- **Trimmed Mean (Moyenne tronquée)** : La moyenne est calculée après avoir éliminé un certain pourcentage des supports les plus élevés et les plus bas pour éviter l'influence des valeurs aberrantes.
- **Minimum/Maximum/Median Rule (Règle du minimum/maximum/médiane)** : Le support total est le minimum, le maximum ou la médiane des supports de tous les classifieurs.
- **Product Rule (Règle du produit)** : Le support total est le produit des supports de tous les classifieurs.

$$\mu_c(x) = \frac{1}{T}\prod_{t=1}^{T} d_{t,c}(x)$$

- **Generalized Mean (Moyenne généralisée)** : Une formule générale qui englobe plusieurs des règles ci-dessus comme cas particuliers, en fonction de la valeur du paramètre α.

$$\mu_c(x) = \left(\frac{1}{T}\sum_{t=1}^{T}(d_{t,c}(x))^{\alpha}\right)^{1/\alpha}$$

Ces méthodes offrent une gamme d'options pour combiner les sorties de classifieurs individuels dans un ensemble, permettant la flexibilité et l'adaptabilité à différents scénarios de problèmes.

2.3.4 Algorithmes populaires basés sur l'ensemble learning

Une riche collection de classificateurs basés sur des ensembles a été développée au cours des dernières années. Toutefois, nombre d'entre eux sont des variantes des quelques algorithmes bien établis dont les capacités ont également été largement testées et rapportées. Dans cette section, nous présentons une vue d'ensemble de certains des algorithmes d'ensemble les plus importants

2.3.4.1 Bagging

L'algorithme de **bagging** (Bootstrap Aggregation) de Breiman est une méthode d'**ensemble learning** simple mais efficace. Il consiste à entraîner plusieurs classificateurs sur différents échantillons **bootstrapés** du jeu de données d'entraînement. La **diversité** dans l'ensemble est assurée par les variations dans ces échantillons et l'utilisation de **classificateurs faibles** comme les *decision stumps* ou les *SVM linéaires*. La décision finale est prise par un **vote majoritaire**. Le **bagging** est particulièrement utile pour les petits jeux de données d'entraînement, tandis qu'une variante appelée *Pasting Small Votes* traite les grands jeux de données en les partitionnant en segments plus petits. L'algorithme de **Random Forest** est une extension notable du **bagging**, com-

binant des **arbres de décision** entraînés avec une sélection aléatoire des instances et des caractéristiques [15].

Algorithm 1 Bagging

Entrées : Données d'entraînement S ; algorithme d'apprentissage supervisé *BaseClassifier* ; entier T spécifiant la taille de l'ensemble ; pourcentage R pour créer les données d'entraînement bootstrapées.

for $t = 1, \ldots, T$ **do**
  Prendre une réplique bootstrapée S_t en sélectionnant aléatoirement $R\%$ de S.
  Appeler *BaseClassifier* avec S_t et recevoir l'hypothèse (classifieur) h_t.
  Ajouter h_t à l'ensemble, $\mathcal{H} = \mathcal{H} \cup \{h_t\}$.
end forCombinaison de l'ensemble : Vote majoritaire simple

Étant donné une instance non étiquetée x :
for $t = 1, \ldots, T$ **do**
  Évaluer l'ensemble $\mathcal{H} = \{h_1, \ldots, h_T\}$ sur x.
  Laisser $v_{t,c} = 1$ si h_t choisit la classe ω_c, et 0 sinon.
end for
Obtenir le total des votes reçus par chaque classe : $V_c = \sum_{t=1}^{T} v_{t,c},\ c = 1, \ldots, C$.
Sortie : Classe avec le plus grand V_c.

2.3.4.2 Boosting and AdaBoost

Boosting, introduit dans le travail *séminal de Schapire* sur le renforcement de l'apprentissage faible, est une approche itérative pour générer un classifieur fort, capable d'atteindre une erreur d'apprentissage arbitrairement faible, à partir d'un ensemble de classifieurs faibles, chacun étant à peine meilleur qu'un choix aléatoire. Bien que le **Boosting** combine également un ensemble de classifieurs faibles en utilisant un **vote majoritaire simple**, il diffère du **Bagging** sur un point crucial. Dans le **Bagging**, les instances sélectionnées pour entraîner des classifieurs individuels sont des répliques bootstrapées des données d'entraînement, ce qui signifie que chaque instance a une chance égale d'être dans chaque ensemble de données d'entraînement. En revanche, dans le **Boosting**, l'ensemble de données d'entraînement pour chaque classifieur suivant se concentre de plus en plus sur les instances mal classées par les classifieurs générés précédemment.

Le **Boosting**, conçu pour des problèmes de classification binaire, crée des ensembles de trois classifieurs faibles à la fois : le premier classifieur (ou hypothèse) h_1 est entraîné sur un sous-ensemble aléatoire des données d'entraînement disponibles, similaire au **Bagging**. Le deuxième classifieur, h_2, est entraîné sur un autre sous-ensemble des données d'origine, dont précisément la moitié est correctement identifiée par h_1 et l'autre moitié est mal classée. Un tel sous-ensemble d'entraînement est considéré comme le "plus informatif", étant donné la décision de h_1. Le troisième classifieur h_3 est ensuite entraîné avec les instances sur lesquelles h_1 et h_2 ne sont pas d'accord. Ces trois classifieurs sont ensuite combinés par un vote majoritaire à trois voies. Schapire a prouvé que l'erreur d'apprentissage de cet ensemble de trois classifieurs est bornée par $g(\epsilon) < 3\epsilon^2 - 2\epsilon^3$, où ϵ est l'erreur de chacun des trois classifieurs, à condition que chaque classifieur ait un taux d'erreur $\epsilon < 0.5$, ce qui est le minimum attendu d'un classifieur dans un problème de classification binaire.

AdaBoost (pour *Adaptive Boosting*) et ses nombreuses variations ont ensuite étendu l'algorithme original de **Boosting** à des problèmes de classification multiple (**AdaBoost.M1**, **AdaBoost.M2**), ainsi qu'à des problèmes de régression (**AdaBoost.R**). Ici, nous décrivons l'**AdaBoost.M1**, la version la plus populaire des algorithmes **AdaBoost**.

L'**AdaBoost** présente deux différences fondamentales par rapport au **Boosting** : (1) les instances sont sélectionnées dans les ensembles de données successifs à partir d'une distribution d'échantillons itérativement

mise à jour des données d'entraînement ; et (2) les classifieurs sont combinés par un **vote majoritaire pondéré**, où les poids des votes sont basés sur les erreurs d'entraînement des classifieurs, elles-mêmes pondérées selon la distribution d'échantillons. La distribution d'échantillons garantit que les échantillons plus difficiles, c'est-à-dire les instances mal classées par le classifieur précédent, sont plus susceptibles d'être inclus dans les données d'entraînement du classifieur suivant.

Le pseudocode de l'**AdaBoost.M1** est fourni dans l'**Algorithme 2**. La distribution d'échantillons, $D_t(i)$, attribue essentiellement un poids à chaque instance d'entraînement x_i, $i = 1, \ldots, N$, à partir de laquelle les sous-ensembles de données d'entraînement S_t sont tirés pour chaque classifieur consécutif (hypothèse) h_t. La distribution est initialisée pour être uniforme ; ainsi, toutes les instances ont la même probabilité d'être incluses dans le premier ensemble de données d'entraînement. L'erreur d'entraînement ϵ_t du classifieur h_t est ensuite calculée comme la somme des poids de distribution de ces instances mal classées par h_t. L'**AdaBoost.M1** exige que cette erreur soit inférieure à 1/2, laquelle est ensuite normalisée pour obtenir β_t, tel que $0 < \beta_t < 1$ pour $0 < \epsilon_t < 1/2$ [15].

Algorithm 2 AdaBoost.M1

Entrées : Données d'apprentissage $\mathcal{D} = \{(x_i, y_i)\}$, $i = 1, \ldots, N$, $y_i \in \{\omega_1, \ldots, \omega_C\}$, apprenant supervisé *BaseClassifier* ; taille de l'ensemble T.
Initialiser : $D_1(i) = \frac{1}{N}, \forall i = 1, \ldots, N$
Pour $t = 1, 2, \ldots, T$ **faire** :

1: Tirer un sous-ensemble d'apprentissage S_t à partir de la distribution D_t.
2: Entraîner *BaseClassifier* sur S_t, recevoir l'hypothèse $h_t : X \to Y$.
3: Calculer l'erreur de h_t :

$$\epsilon_t = \sum_{i=1}^{N} \mathbf{I}[h_t(x_i) \neq y_i] \cdot D_t(i)$$

4: **if** $\epsilon_t > \frac{1}{2}$ **then**
5: **abandonner**.
6: **end if**
7: Définir

$$\beta_t = \frac{\epsilon_t}{1 - \epsilon_t}$$

8: Mettre à jour la distribution d'échantillonnage :

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times \begin{cases} \beta_t, & \text{si } h_t(x_i) = y_i \\ 1, & \text{sinon} \end{cases}$$

où $Z_t = \sum_{i=1}^{N} D_t(i)$ est une constante de normalisation pour garantir que D_{t+1} est une fonction de distribution appropriée.
9: **Vote majoritaire pondéré :** Étant donné une instance non étiquetée z, obtenir le total des votes reçus par chaque classe :

$$V_c = \sum_{t:h_t(z)=\omega_c} \log \frac{1}{\beta_t}, \quad c = 1, \ldots, C$$

Sortie : Classe avec le plus grand V_c.

Parmi les travaux internationaux sur le chikungunya en utilisant l'ensemble learning nous avons :

Data Science System of Predicting and Detecting the Chikungunya Virus using Ensemble Techniques par **Bommireddy Vijay Kumar Reddy, Shail Patel, Anayna Singh** et **al** le *29 Decembre 2022* : . L'objectif principal des auteurs est d'utiliser des techniques d'ensemble pour trouver une approche novatrice et comparer ces techniques d'ensemble.

Chapitre 3

Implémentation des Modèles

3.1 Introduction

3.2 Présentation des Outils Utilisés

Bla

3.2.1 Langages et librairies

Bla

3.2.2 Infrastructure matérielle

Bla

3.3 Méthodes de l'Apprentissage Automatique

Bla

3.3.1 Le jeu de données

Bla

3.3.2 Préparation et nettoyage des données

jijj

3.3.3 Implémentation des modèles de machine learning

Bla

3.4 Méthodes de l'Apprentissage Profond

Bla

3.4.1 Le jeu de données

Bla

3.4.2 Préparation des données pour le deep learning

Bla

3.4.3 Implémentation des modèles de deep learning

Bla

Conclusion

General Conclusion

Bibliographie

[1] Ferreira de Almeida, I., Codeço, C.T., Lana, R.M., Bastos, L.S., de Souza Oliveira, S., Andreza da Cruz Ferreira, D., Godinho, V.B., Souza Riback, T.I., Cruz, O.G., Coelho, F.C. : The expansion of chikungunya in brazil. Lancet Reg. Health Am. **25**(100571), 100,571 (2023)

[2] Auerswald, H., Boussioux, C., In, S., et al. : Broad and long-lasting immune protection against various chikungunya genotypes demonstrated by participants in a cross-sectional study in a cambodian rural community. Emerging Microbes & Infections **7**(1), 13 (2018). DOI 10.1038/s41426-018-0013-0

[3] CDC : Transmission of chikungunya virus. `https://www.cdc.gov/ncezid` Visité le 10/juin/2024

[4] cdc — US centers for disease control, prevention : National center for emerging and zoonotic infectious diseases (ncezid). `https://www.cdc.gov/ncezid` Visité le 12/juin/2024

[5] de Souza, W.M., Ribeiro, G.S., de Lima, S.T., de Jesus, R., Moreira, F.R., Whittaker, C., Sallum, M.A.M., Carrington, C.V., Sabino, E.C., Kitron, U., Faria, N.R., Weaver, S.C. : Chikungunya : a decade of burden in the americas. The Lancet Regional Health - Americas **30**, 100,673 (2024). DOI https://doi.org/10.1016/j.lana.2023.100673. URL `https://www.sciencedirect.com/science/article/pii/S2667193X23002478`

[6] Dr Jean Bosco NDIHOKUBWAYO, D.B.H.e.a. : Rapport de la situation Épidémiologique chikungunya

[7] ECDC : Chikungunya worldwide overview. `https://www.ecdc.europa.eu/en/chikungunya-monthly` Visité le 10/juin/2024

[8] Ganesan, V., Duan, B., Reid, S. : Chikungunya virus : Pathophysiology, mechanism, and modeling. Viruses **9**(12), 368 (2017). DOI 10.3390/v9120368

[9] Giovanetti, M., Vazquez, C., Lima, M., Castro, E., Rojas, A., de la Fuente, A.G., Aquino, C., Cantero, C., Fleitas, F., Torales, J., et al. : Rapid epidemic expansion of chikungunya virus east/central/south african lineage, paraguay. Emerging Infectious Diseases **29**(9), 1859 (2023)

[10] Mahesh, B. : Machine learning algorithms - a review **9**. DOI 10.21275/ART20203995

[11] Mehrbakhsh Nilashi Othman bin Ibrahim, H.A.L.S. : An analytical method for diseases prediction using machine learning techniques DOI http://dx.doi.org/doi:10.1016/j.compchemeng.2017.06.011

[12] Moreira, J., Soares, C., Jorge, A., Sousa, J. : Ensemble approaches for regression : A survey. ACM Computing Surveys **45**, 10 :1–10 :40 (2012). DOI 10.1145/2379776.2379786

[13] Morrison, T.E. : Reemergence of chikungunya **88**(20), 11,644 – 11,647. DOI https://doi.org/10.1128/jvi.01432-14

[14] PAHO : Chikungunya - paho/who — pan american health organization. `https://www.paho.org/en/topics/chikungunya` Visité le 12/juin/2024

[15] Polikar, R. : Ensemble learning. Ensemble machine learning : Methods and applications pp. 1–34 (2012)

[16] Reddy, B.V.K., Patel, S., Singh, A., Singh, N., Ranjit, S., Dantkale, S. : Data science system of predicting and detecting the chikungunya virus using ensemble techniques. In : 2022 4th International Conference on Inventive Research in Computing Applications (ICIRCA), pp. 1420–1423 (2022). DOI 10.1109/ICIRCA54612.2022.9985708

[17] Rougeron, V., Sam, I.C., Caron, M., Nkoghe, D., Leroy, E., Roques, P. : Chikungunya, a paradigm of neglected tropical disease that emerged to be a new health global risk. Journal of clinical Virology **64**, 144–152 (2015)

[18] WHO : Chikungunya — world health organization. `https://www.who.int/health-topics/chikungunya#tab=tab_1` Visité le 10/juin/2024

[19] WHO : Chikungunya fact sheet. `https://www.who.int/news-room/fact-sheets/detail` Visité le 12/juin/2024

[20] WHO : Disease outbreak news - tchad. `https://www.who.int/fr/emergencies/disease-outbreak-news/item/chikungunya-chad` Visité le 10/juin/2024