Epi_DCA : Adaptation et mise en œuvre de la théorie du danger pour la veille épidémiologique

Bahdja Boudoua[1,3], Mathieu Roche[1,4],
Maguelonne Teisseire [1,3], Annelise Tran[1,2,4]

[1] UMR TETIS, Univ. Montpellier, AgroParisTech, CIRAD, CNRS, INRAE, Montpellier, France.
[2]UMR ASTRE, Univ. Montpellier, CIRAD, INRAE, Montpellier, France.
[3]INRAE, UMR TETIS, Montpellier, France.
[4] CIRAD, UMR TETIS, F-34398 Montpellier, France.

Résumé. Le rôle des systèmes de surveillance basés sur les évènements (SBE) est de détecter les nouvelles épidémies en explorant les informations sanitaires publiées en ligne dans un large éventail de sources formelles et informelles. Les facteurs de risque (environnementaux, climatiques, liés aux pratiques d'élevage etc.) ne sont en général pas pris en compte par ces systèmes. Dans cet article, nous souhaitons poser les premières bases d'une démarche générique (indépendante d'une maladie ou d'un hôte spécifique) qui permet de renforcer ou non un événement détecté par les systèmes de veille en y intégrant les facteurs de risque disponibles. Epi_DCA est une adaptation de l'algorithme des cellules dendritiques (DCA), inspiré de la théorie du danger. Il permet de combiner les différents facteurs de risque aux données épidémiologiques issues des systèmes de veille. Un premier test est effectué sur le cas d'étude Influenza aviaire (IA). Par la suite, l'approche proposée sera évaluée sur différents cas d'étude (fièvre du Nil occidental et peste porcine africaine) afin de tester sa robustesse et sa généricité.

1 Introduction

Pour faire face aux maladies émergentes qui représentent un risque croissant pour la santé publique, de nombreux pays adoptent une stratégie de veille sanitaire. Celle-ci repose sur deux composantes : la surveillance basée sur les indicateurs (SBI) (issus de sources officielles telles que l'OIE (Office International des Epizooties : Organisation mondiale de la santé animale), l'OMS (Organisation Mondiale de la Santé) ou la FAO (Food and Agriculture Organization)) et la surveillance basée sur les évènements (SBE) issus de sources non-officielles (médias en ligne, réseaux sociaux, etc.).
Les systèmes de SBE tels que ProMed (Carrion and Madoff, 2017), HealthMap (Freifeld et al., 2008), et PADI-Web (Valentin et al., 2020) sont utilisés quotidiennement afin de détecter les évènements de santé inhabituels. Ils collectent et analysent un flux quotidien de données textuelles non structurées (articles) à partir d'internet, en utilisant des mots-clés ou des combinaisons de mots-clés (Barboza et al., 2014). Par la suite, ces articles sont triés selon leur pertinence

et classés par date, localisation géographique, source, et maladie. La surveillance basée sur les évènements permet ainsi l'obtention de nombreuses informations mais présente certaines limites. En particulier, les facteurs de risque liés à l'apparition des maladies ne sont pas toujours retrouvés dans les données textuelles et ne sont pas pris en compte par les SBE. Parallèlement, la cartographie du risque en épidémiologie permet de mettre en évidence les zones favorables à l'apparition d'une maladie en s'appuyant sur la répartition spatiale des facteurs de risque associés (Hess et al., 2018). La connaissance de ces facteurs de risque est essentielle pour mieux cibler les zones de surveillance et adapter les mesures de lutte et de prévention (Bergmann et al., 2021). Dans cet article, les premières bases d'une approche inspirée de la théorie du danger sont établies : Epi_DCA est l'adaptation de l'algorithme des cellules dendritiques (DCA) à la problématique de veille sanitaire et afin de classer les articles analysés comme pertinents (l'article traite d'une émergence ou ré-émergence d'influenza aviaire) vs non-pertinents. Cette méthode permet de combiner les facteurs de risque aux données épidémiologiques issues des systèmes de veille tout en prenant en compte la dimension spatio-temporelle des évènements épidémiologiques.

La suite de l'article est structurée de la façon suivante. L'état de l'art sur la théorie du danger et les travaux relatifs au DCA sont présentés en Section 2. Notre méthode Epi_DCA est décrite en Section 3 puis un cas d'étude et les résultats préliminaires sont présentés en Section 4. Un bilan et les perspectives sont abordés en Section 5.

2 État de l'art

2.1 Théorie du danger

La théorie du danger (Matzinger, 2002) est basée sur le fonctionnement des cellules immunitaires dendritiques (DCs). Les DCs jouent un rôle essentiel dans le déclenchement des réponses immunitaires. Elles sont parmi les premières cellules exposées à l'environnement extérieur et ont la capacité de détecter et d'interpréter une multitude d'informations moléculaires potentiellement contradictoires. L'interprétation de ces informations (signaux) au système immunitaire adaptatif conduit au déclenchement ou non d'une réponse contre les menaces perçues.

La théorie du danger stipule que la reconnaissance d'un antigène par une DC ne réside pas dans la distinction entre le soi et le non-soi mais dépend plutôt du contexte environnemental (signaux) dans lequel l'antigène est identifié. Les DCs existent dans l'un des trois états suivants : "immature", "semi-mature" et "mature". A leur état initial, les DCs sont "immatures". Ensuite, en fonction de la concentration des signaux auxquels elles sont exposées, elles se différencient soit en cellules "semi-matures" pour inhiber la réponse immunitaire, soit en cellules "matures" pour l'activer (Gu et al., 2013).

Cette théorie et le comportement des DCs a servi d'inspiration pour le développement d'un algorithme de classification, l'algorithme des cellules dendritiques (DCA) (Greensmith, 2007).

Le DCA a été appliqué avec succès à un large éventail d'applications, par exemple dans le domaine de la sécurité informatique (Jim et al., 2022; Sharaff et al., 2021), ou encore en sismologie (Zhou et al., 2020). Il présente les avantages suivants lorsqu'il est appliqué à des problèmes en temps réel :

— Il ne nécessite pas de longues périodes d'apprentissage ;

— Il permet d'intégrer des données hétérogènes par le biais de deux types de signaux ;
— Il a montré des résultats prometteurs quant à la réduction du nombre de faux positifs (Mohsin et al., 2014).

2.2 Algorithme des cellules dendritiques (DCA) et travaux associés

Le DCA a été initialement conçu pour être utilisé comme un algorithme de détection d'anomalies.
Son processus comprend 4 phases : pré-traitement et catégorisation des données, détection des antigènes (AGs) par les cellules, évaluation du contexte cellulaire, et classification des antigènes.
La première phase comprend deux étapes principales : la réduction des attributs et la catégorisation du signal. Pour réduire le nombre d'attributs, certains travaux ont recours à l'avis d'experts lors de la phase de pré-traitement des données. D'autres utilisent des méthodes statistiques telles que l'analyse en composantes principales (ACP) (Chelly and Elouedi, 2016). Les caractéristiques les plus intéressantes de l'ensemble de données sont ainsi sélectionnées puis classées dans l'une des catégories de signaux définies du DCA, à savoir :

— Les signaux de danger (*danger signals*) qui augmentent proportionnellement à la présence de données représentant une situation "anormale" ;
— Les signaux sécuritaires (*safe signals*) qui augmentent proportionnellement à la présence de données représentant une situation "normale".

Après cette première phase, chaque antigène est caractérisé par son signal de danger et son signal sécuritaire.
Lors de la phase de détection, chaque cellule dendritique (DC) est exposée aléatoirement à des antigènes (AGs). Les signaux de sortie cumulés (CSM) sont calculés selon l'équation suivante :

$$Eq.1 \qquad CSM = W_{SS} \times S_S + W_{DS} \times S_D$$

où S_S et S_D représentent les valeurs et W_{SS} et W_{DS} les poids des signaux sécuritaires (S) et des signaux de danger (D), respectivement.
Les pondérations utilisées dans le DCA peuvent être dérivées empiriquement des données ou de valeurs définies par l'utilisateur. Les signaux de danger ont toujours un poids positif et les signaux sécuritaires un poids négatif.
Les CSM des DCs ont deux rôles : d'abord de définir un contexte cellulaire (mature ou semi-mature), et ensuite de stopper l'exposition des cellules aux antigènes (Farzadnia et al., 2021). En effet, pour limiter le temps d'exposition des DCs, chaque DC se voit attribuer une valeur de seuil de migration lors de sa création. Suite à la mise à jour des CSM, si la valeur de CSM dépasse la valeur du seuil de migration, alors l'exposition est stoppée (Chelly Dagdia and Elouedi, 2020).
Lors de la troisième phase, le contexte cellulaire est utilisé pour étiqueter les antigènes collectés par les DCs, et cette information est finalement utilisée dans la génération d'un coefficient d'anomalie qui sera traité dans la phase finale (classification) (Chelly and Elouedi, 2016). Le coefficient d'anomalie noté MCAV (pour Moleculare Antigen Value) reflète le degré d'anomalie d'un antigène donné (plus le MCAV est proche de 1, plus la probabilité qu'un antigène soit anormal est grande), et est calculé en divisant le nombre de DCs matures par le nombre total

de DCs exposées à un antigène, comme indiqué dans l'Eq.2

$$Eq.2 \qquad \text{MCAV} = \frac{\alpha}{(\alpha+\beta)}$$

où α et β sont respectivement le nombre de DC matures et immatures qui ont été exposées à l'antigène.
Une fois le MCAV calculé pour chaque antigène, dans sa dernière phase l'algorithme DCA peut effectuer sa tâche de classification, en comparant le MCAV de chaque antigène à un seuil d'anomalie (paramètre défini par l'utilisateur, ou estimé par apprentissage).
Les premières versions du DCA présentaient un nombre important de paramètres et d'éléments stochastiques (Greensmith, 2007), tels que l'exposition aléatoire et les seuils de migration variables des cellules (Greensmith and Aickelin, 2008). L'estimation de ces paramètres était souvent faite de façon arbitraire et cette limite de l'algorithme a été soulignée par plusieurs études critiques (Chelly and Elouedi, 2016). Différentes versions ont déjà été présentées afin de réviser et d'améliorer le DCA en fonction du problème étudié (Elisa et al., 2018; Zhou and Liang, 2021).
Cependant, il n'y a pas suffisamment d'études qui abordent la problématique de l'exposition aléatoire et de la définition du seuil de migration des DCs. D'autre part, à notre connaissance le DCA n'a pas été appliqué à la thématique de la veille sanitaire de manière à intégrer données épidémiologiques et leurs facteurs de risque en prenant en compte leurs dimensions spatiale et temporelle.
Les contributions principales du travail présenté ici sont les suivantes :

1. Nous adaptons le DCA à la problématique de la veille sanitaire afin de combiner les différents facteurs de risques aux données épidémiologiques issues des SBE ;
2. Nous intégrons l'information spatiale dans la phase de détection. L'exposition des cellules n'est pas aléatoire mais dépend de la distance entre les DCs et les AGs, ainsi que recommandé par (Farzadnia et al., 2021) ;
3. Enfin, nous intégrons l'information temporelle afin d'obtenir un seuil de migration indépendant de la valeur des signaux cumulés CSM.

3 Transposition à la veille sanitaire

3.1 Pré-traitement et catégorisation des données

Dans le contexte de nos travaux, les évènements extraits d'articles détectés par les systèmes SBE représentent nos antigènes (ce que l'on veut classer), associés à des données environnementales (Figure 1 étape de collecte des données) par une correspondance spatiale. Dans une phase de pré-traitement et catégorisation (Figure 1 Phase 1), ces données d'entrées sont converties en deux catégories de signaux : signaux de danger (*danger signals*) et signaux sécuritaires (*safe signals*). Les données épidémiologiques issues de ces articles (source de l'information, hôte, maladie) constituent les signaux de danger.

Nous nous référons à la connaissance d'experts afin d'établir un barème et donner une note à chaque donnée observée, par exemple : +20 si la source de l'article est officielle, +15 si l'hôte est un animal domestique, etc. Si aucune de ces données n'est présente, le signal de danger est

nul et l'article n'est pas pris en compte.
Les données environnementales représentent les signaux sécuritaires. Un signal sécuritaire maximal indique que l'environnement est défavorable à l'apparition de la maladie, un signal sécuritaire égal à 0 indique au contraire un environnement favorable dans lequel tous les facteurs de risque identifiés sont présents.

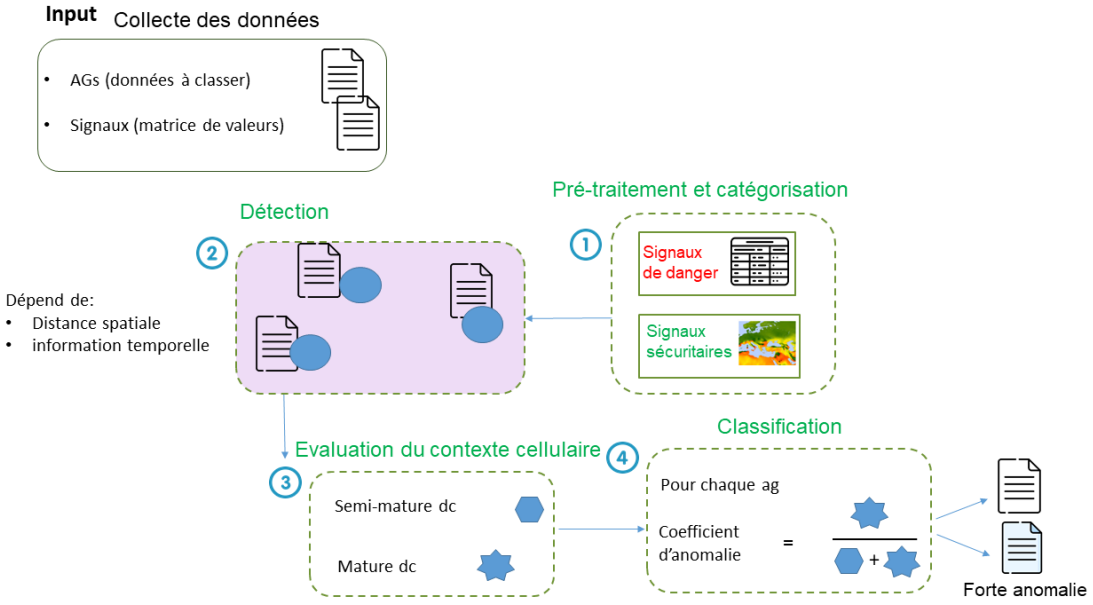

FIG. 1 – *Epi_DCA illustré en 4 phases*

À la suite du pré-traitement et catégorisation des données, Epi_DCA s'articule en 3 phases.

3.2 Phase de détection

A la phase de détection, les DCs sont exposées aux AGs (Figure 1 Phase 2). Le CSM est ensuite calculé en combinant les signaux pondérés de danger et sécuritaires. La pondération est adaptée à la maladie étudiée : pour une maladie fortement influencée par les facteurs environnementaux (par exemple une maladie à transmission vectorielle comme la fièvre du Nil Occidental, ou impliquant un réservoir sauvage, comme l'influenza aviaire) un poids plus important sera attribué aux signaux sécuritaires qui caractérisent le contexte environnemental dans notre approche.
Dans Epi_DCA, l'exposition des DCs dépend de : (1) la distance spatiale entre les DCs et les AGs et (2) le rayon de couverture R des DCs (Figure 2). À chaque pas de temps, les signaux

de sorties cumulés (CSM) des DCs sont mis à jour selon :

$$\begin{cases} CSM_{t+1} = CSM_t + (\Delta_{dist} \times CSM_{entrant}) \\ CSM_0 = 0 \end{cases}$$

avec Δ_{dist} un coefficient de distance inversement proportionnel à la distance spatiale. En d'autres termes, plus la distance est grande, plus la contribution du $CSM\_entrant$ est faible. Cela traduit le fait que la propagation de certaines maladies est liée à la distance entre les évènements observés (Salje et al., 2016).

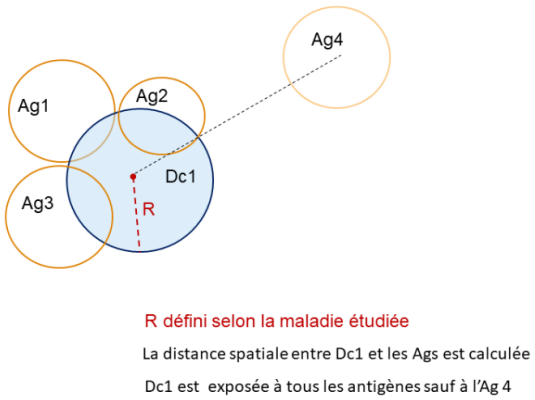

FIG. 2 – *Exemple d'exposition d'une cellule immunitaire dendritique (DC) aux antigènes (AGs)*

Selon la littérature, le seuil de migration (mt) des DCs est déterminé par l'utilisateur, et l'exposition d'une DC aux AGs ne s'arrête que si la valeur cumulée du signal de sortie dépasse le mt. Dans Epi_DCA la migration des DCs dépend aussi de l'information temporelle des antigènes entrants (autrement dit la date de publication des articles) : l'exposition des DCs est interrompue et les signaux cumulés sont remis à 0 lorsqu'aucun antigène n'est détecté pendant une période de temps définie (en fonction de la maladie considérée), ce qui permet de ne pas biaiser la surveillance par des évènements antérieurs (lointains) au même emplacement qui n'ont plus d'impact sur un évènement épidémiologique en cours.
Le pseudo-code complet pour effectuer la phase de détection est présenté dans **Phase 2 - Procedure Detection phase**.

3.3 Phase d'évaluation du contexte cellulaire

La phase d'évaluation du contexte est effectuée une fois que les DCs ont migré, l'évaluation du contexte cellulaire prend en compte les signaux de sortie cumulés et le nombre d'exposition de chaque cellule. Chaque cellule est étiquetée comme "DC mature" ou "DC semi-mature". Le label "mature" signifie que la cellule concernée a été grandement exposée aux signaux de

danger durant la période définie contrairement aux cellules "semi-matures". Cette information sera utilisée lors de la phase de classification.
Le pseudo-code complet permettant de définir le contexte cellulaire est présenté dans **Phase 3 - Procedure Context Assessment phase**.

Phase 2 – Procedure Detection phase

procedure DETECTION PHASE
  Input $ag : antigen\ (Date,\ Coord,\ Host,\ Source,\ Subtype, \emptyset)$
  $DCs :\ Cell\ (Date,\ Coord,\ NbExp,\ CSM,\ Context)$
  Output updated ag : antigen (Date, Coord, Host, Source, Subtype, ListCell), updated DCs : Cell (Date, Coord, NbExp ,CSM, Context)
  ▷ In the detection phase, all the DCs are exposed to the incoming antigen ag.
  ▷ Compute the exposure for each cell to the antigen (if exists)
  for each cell dc of DCs **do**
    $DistCellAnt \leftarrow calculation\ distance\ between\ ag\ and\ dc$
    $DiffdaysCellAnt \leftarrow calculation\ diffdays\ between\ ag\ and\ dc$
    if $DistCellAnt < Disease.space\_limit$ **and**
      $DiffdaysCellAnt < Disease.time\_limit$ **then**
        $\Delta_{dist} \leftarrow \frac{(Disease.space\_limit - DistCellAnt)}{Disease.space\_limit}$
        $dc.CSM \leftarrow dc.CSM + (\Delta\_dist \times dc.CSM)$
        $dc.NbExp \leftarrow dc.NbExp + 1$
        $ag.ListCell \leftarrow ag.ListCell + dc$
    end if
  end for
  $DCs \leftarrow DCs + NewCell(ag, Disease)$
end procedure

Phase 3 – Procedure Context Assessment Phase

Input DCs (Date, Coord, NbExp, CSM, Context)
Output updated : DCs (Date, Coord, NbExp, CSM, Context)

$Ratio\_Exp = \frac{Mean(DCs.CSM)}{Mean(DCs.NbExp)}$

▷ $Ratio\_Exp$ depends on the disease, it is used as a threshold to assign the dc.context

for each cell dc of DCs **do**
  if $\frac{dc.CSM}{dc.NbExp} > Ratio\_Exp$ **then**
    $dc.context \leftarrow mature$
  else
    $dc.context \leftarrow semi - mature$
  end if
end for

▷ the dc.context will be used in the last phase (classification) to generate an anomaly coefficient for each ag

3.4 Phase de classification

Enfin, dans la phase de classification des évènements détectés par les SBE (Figure 1 phase 4), les signaux sortants sont utilisés pour générer un coefficient d'anomalie propre à chaque antigène et qui prend ainsi en compte à la fois les informations sanitaires issues des articles (danger signals), le contexte environnemental (safe signals), et l'information spatio-temporelle des évènements épidémiologiques.
Ce coefficient d'anomalie est compris entre 0 et 1, plus sa valeur tend vers 1 plus la probabilité que l'antigène soit anormal est grande. Le seuil d'anomalie (AnomalyThreshold) est fixé à 0.5 comme cela est proposé dans la littérature (Chelly and Elouedi, 2016). Le pseudo-code complet pour la phase de classification est présenté dans **Phase 4 - Procedure Classification phase**.

Phase 4 – Procedure Classification phase

Input AGs set of antigens
Output updated : CoeffAnomaly of each ag of AGs

for each antigen ag of AGs **do**
Compute anomaly coefficient
  $Coeff \leftarrow \frac{ag.ListCell.mature()}{ag.ListCell.length()}$
  ▷ sum of matures cells exposed to ag divided by sum of total exposed cells to ag
  if $Coeff > disease.AnomalyThreshold$ **then**
    $ag.AnomalyCoef \leftarrow anomalous$
  else
    $ag.AnomalyCoef \leftarrow normal$
  end if
end for

4 Cas d'étude IA et résultats préliminaires

Le premier cas d'étude auquel nous nous intéressons est la grippe aviaire (IA - influenza aviaire), d'une part parce qu'il s'agit d'une maladie médiatisée et de ce fait, un nombre conséquent d'articles est détecté par les SBE et d'autre part parce que l'émergence et la diffusion de cette maladie dépendent de différents facteurs de risque (proximité des zones humides, population d'oiseaux sauvages, population d'oiseaux domestiques, etc.). Pour des raisons de disponibilité des données, nous nous sommes intéressés à la région d'Asie du Sud-Est et à la base de données HealthMap (Freifeld et al., 2008).

4.1 Collecte des données et paramètres

Nous avons constitué un jeu de données de 174 articles publiés entre août 2018 et juillet 2019, issus du système de veille HealthMap. Le jeu de données, qui sera rendu public, est constitué actuellement de 87 articles pertinents et 87 articles non pertinents. Dans ce contexte, un article pertinent décrit au minimum un évènement (foyer) d'IA avec sa localisation spatiale (figure 3). Un article non pertinent décrit des mesures sanitaires/économiques, ou traite d'une maladie différente de l'influenza aviaire. Ces articles ont été classés selon leur pertinence manuellement après lecture des textes par un épidémiologiste.

BEIJING (Reuters) - China has confirmed two cases of H5N6 avian bird flu on poultry farms in southwestern province of Yunnan, the Agriculture Ministry said on Wednesday.

Local authorities have culled 10,280 birds following the outbreaks, the Ministry of Agriculture and Rural Affairs said in a statement on its website.

Outbreaks infected a total of 11,340 birds in two farms in Tengchong city and Luquan county in Yunnan, and killed 9,820 of them, the statement said.

FIG. 3 – *Extrait d'un article pertinent détecté par Healthmap*

4.2 Signaux de danger

Les données épidémiologiques issues des articles détectés (source d'information, hôte, maladie) sont utilisées pour générer les signaux de danger. Nous nous référons aux connaissances d'experts afin d'établir un score pour chaque donnée observée **(Tableaux 1 et 2)**. Par exemple, les sources officielles telles que l'OIE, la FAO ont un score plus élevé que les sources non officielles (médias en ligne, réseaux sociaux) et l'influenza aviaire hautement pathogène (IAHP) a un score plus élevé que l'influenza aviaire faiblement pathogène (LPAI). Les signaux de danger ont par la suite été affinés de façon empirique.

Articles ID	**Epidemiological data**		
	Source	**Subtype**	**Host**
ID_1	FAO	HPAI	Wild birds
ID_2	Twitter	Unspecified	Unspecified
ID_3	OIE	LPAI	Domestic birds

TAB. 1 – *Aperçu de la base de données Healthmap*

Articles	**Danger signals**			
	Source	**Subtype**	**Host**	**Total**
Ag_1	30	40	30	**100**
Ag_2	20	0	10	**30**
Ag_2	30	30	20	**80**

TAB. 2 – *Données épidémiologiques (Tableau 1) converties en signaux de danger*

4.3 Signaux sécuritaires

Pour générer les signaux sécuritaires, nous avons créé une carte de risque d'occurrence d'IA selon la méthode décrite par (Stevens et al., 2013) en utilisant des données récentes sur les populations humaines, les oiseaux domestiques et sauvages (figure 4). Ensuite, les événements ont été associés aux données environnementales par correspondance spatiale à l'aide d'un Système d'Information Géographique (SIG). Le logiciel QGIS [1] a été utilisé.

La valeur des signaux sécuritaires est comprise entre 0 et 100 et diminue proportionnellement à la valeur de l'indice de risque d'occurrence d'IA.

4.4 Rayon de couverture et seuil de migration des DCs

Le virus de l'IA est susceptible de se propager par différentes voies (transport de volailles, migration des oiseaux...etc) (Yousefinaghani et al., 2020). Cela rend sa dynamique de diffusion complexe et difficile à déterminer. Ici nous fixons le rayon de couverture des DCs à 20 km, qui correspond à la distance pour laquelle les restrictions et mesures de contrôle sont mises en place (zone de surveillance) autour des foyers d'IA (Pittman and Laddomada, 2008). Quant au seuil de migration des DCs, il a été fixé à 21 jours, car au-delà de cette période si aucun nouvel évènement d'IA n'est détecté, le foyer concerné est considéré comme assaini.

4.5 Résultats

Les résultats de la classification des articles appartiennent à l'une des classes suivantes : True Positive (TP) : Article pertinent correctement classé, True Negative (TN) : Article non-pertinent correctement classé, False Positive (FP) : Article non-pertinent classé comme pertinent et False Negative (FN) : Article pertinent classé comme non-pertinent. Les métriques

1. www.qgis.org

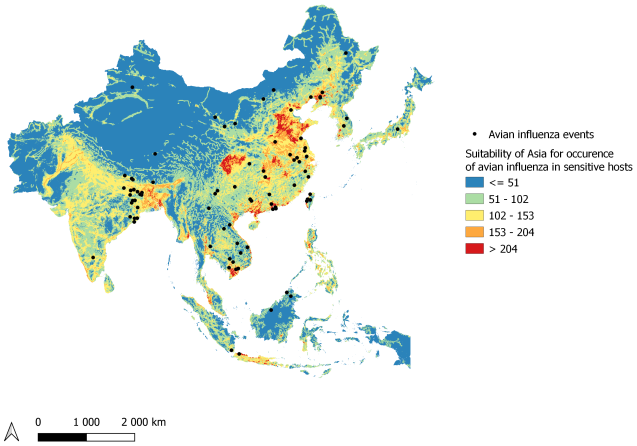

FIG. 4 – *Carte de risque d'occurrence d'influenza aviaire en Asie du Sud-Est obtenue par évaluation multicritère spatialisée (Stevens et al., 2013). L'indice de risque est compris entre 0 (risque faible) et 255 (risque fort)*

utilisées pour évaluer les performances de l'algorithme sont Précision, Rappel, et F-score. Ces métriques sont calculées selon les équations suivantes :

$$Précision = \frac{TP}{(TP + FP)} \tag{1}$$

$$Rappel = \frac{TP}{(TP + FN)} \tag{2}$$

$$F - Score = 2 \text{ X } \frac{Précision \text{ X } Rappel}{Précision \ + \ Rappel} \tag{3}$$

Métrique	Epi_DCA	EPI_DCA sans safe signals
Précision	0.850	0.827
Rappel	0.870	0.782
F-score	0.860	0.80

TAB. 3 – *Résultats de classification - Epi_DCA - 1er cas d'étude avec et sans intégration des signaux sécuritaires*

Dans un premier temps, notre méthode a été testée avec et sans intégration des signaux sécuritaires. Les résultats indiqués dans le Tableau 3, sont encourageants et suggèrent que la prise

en compte du contexte environnemental (dans ses dimensions spatio-temporelles) dans l'analyse des données épidémiologiques issues des SBE permet de renforcer les articles détectés par les SBE.

Par la suite, le logiciel Weka [2] a été utilisé pour tester quatre méthodes d'apprentissage supervisé (SVM, Naive Bayes, Knn et Random Forest) sur notre jeu de données en effectuant une validation croisée en 5 plis. Nous avons obtenu une F-mesure entre 0.861 (Naive Bayes) et 0.913 (SVM) ce qui montre que notre approche Epi_DCA, qui a la caractéristique d'être non supervisée, reste tout à fait compétitive.

Métriques	**Résultats**			
	SVM	Naive Bayes	K-nn	Random Forest
Précision	0.923	0.869	0.882	0.918
Rappel	0.914	0.862	0.879	0.902
F-Score	0.913	0.861	0.879	0.901

TAB. 4 – *Résultats de classification obtenus avec les méthodes d'apprentissage supervisé*

5 Conclusion et perspectives

Dans cette étude nous avons posé les premières bases d'Epi_DCA qui est l'adaptation du DCA à la problématique de la veille sanitaire. Les principales contributions de ce travail sont la prise en compte des facteurs de risque associés à la maladie ciblée ainsi que l'intégration de la dimension spatio-temporelle des évènements épidémiologiques dans la méthode. A partir du pseudo code présenté ici, nous avons implanté Epi_DCA en utilisant le langage R qui donnera lieu à une librairie que nous rendrons disponible pour la communauté. Les paramètres des deux types de signaux permettent d'ajuster la sensibilité de l'algorithme et de l'adapter à la maladie ciblée. Ainsi, en changeant le jeu de données et/ou la maladie cible, il nous est possible de se calibrer simplement et efficacement et d'augmenter grandement la réutilisabilité de notre code. Epi_DCA a été testé et évalué dans un premier temps sur le cas d'étude influenza aviaire et a montré des résultats prometteurs. Cette méthode ne nécessite pas d'apprentissage, elle permet de combiner des données hétérogènes par le biais de deux types de signaux et a montré de bons résultats quant à la réduction du nombre de faux positifs. Ces avantages ont conduit à l'explorer dans le contexte de veille sanitaire. La méthode proposée sera testée sur d'autres cas d'étude pour tester sa généricité : pour une maladie à transmission vectorielle comme la fièvre du Nil Occidental et pour une maladie transfrontalière comme la peste porcine africaine, et ce dans différents contextes géographiques.

Remerciements

Ce travail est financé par le projet « Monitoring outbreak events for disease surveillance in a data science context » (MOOD) du programme de recherche et d'innovation Horizon 2020

2. https ://www.cs.waikato.ac.nz/ml/weka/index.html

de l'Union européenne dans le cadre de la convention de subvention n° 874850 (https ://mood-h2020.eu /). Nous remercions également le projet HealthMap (https ://healthmap.org/), qui nous a fourni les données.

Références

P. Barboza, L. Vaillant, Y. Le Strat, D. M. Hartley, N. P. Nelson, A. Mawudeku, L. C. Madoff, J. P. Linge, N. Collier, J. S. Brownstein, et al. Factors influencing performance of internet-based biosurveillance systems used in epidemic intelligence for early detection of infectious diseases outbreaks. *PloS one*, 9(3) :e90536, 2014.

H. Bergmann, K. Schulz, F. J. Conraths, and C. Sauter-Louis. A review of environmental risk factors for african swine fever in european wild boar. *Animals*, 11(9) :2692, 2021.

M. Carrion and L. C. Madoff. Promed-mail : 22 years of digital surveillance of emerging infectious diseases. *International health*, 9(3) :177–183, 2017.

Z. Chelly and Z. Elouedi. A survey of the dendritic cell algorithm. *Knowledge and Information Systems*, 48(3) :505–535, 2016.

Z. Chelly Dagdia and Z. Elouedi. A hybrid fuzzy maintained classification method based on dendritic cells. *Journal of Classification*, 37(1) :18–41, 2020.

N. Elisa, L. Yang, and N. Naik. Dendritic cell algorithm with optimised parameters using genetic algorithm. In *2018 IEEE Congress on Evolutionary Computation (CEC)*, pages 1–8. IEEE, 2018.

E. Farzadnia, H. Shirazi, and A. Nowroozi. A new intrusion detection system using the improved dendritic cell algorithm. *The Computer Journal*, 64(8) :1193–1214, 2021.

C. C. Freifeld, K. D. Mandl, B. Y. Reis, and J. S. Brownstein. Healthmap : global infectious disease monitoring through automated classification and visualization of internet media reports. *Journal of the American Medical Informatics Association*, 15(2) :150–157, 2008.

J. Greensmith. *The dendritic cell algorithm*. PhD thesis, Citeseer, 2007.

J. Greensmith and U. Aickelin. The deterministic dendritic cell algorithm. In *International conference on artificial immune systems*, pages 291–302. Springer, 2008.

F. Gu, J. Greensmith, and U. Aickelin. Theoretical formulation and analysis of the deterministic dendritic cell algorithm. *Biosystems*, 111(2) :127–135, 2013.

A. Hess, J. Davis, and M. Wimberly. Identifying environmental risk factors and mapping the distribution of west nile virus in an endemic region of north america. *GeoHealth*, 2(12) : 395–409, 2018.

L. E. Jim, N. Islam, and M. A. Gregory. Enhanced manet security using artificial immune system based danger theory to detect selfish nodes. *Computers & Security*, 113 :102538, 2022.

P. Matzinger. The danger model : a renewed sense of self. *science*, 296(5566) :301–305, 2002.

M. F. M. Mohsin, A. A. Bakar, and A. R. Hamdan. Outbreak detection model based on danger theory. *Applied soft computing*, 24 :612–622, 2014.

M. Pittman and A. Laddomada. Legislation for the control of avian influenza in the european union. *Zoonoses and public health*, 55(1) :29–36, 2008.

H. Salje, D. A. Cummings, and J. Lessler. Estimating infectious disease transmission distances using the overall distribution of cases. *Epidemics*, 17 :10–18, 2016.

A. Sharaff, C. Kamal, S. Porwal, S. Bhatia, K. Kaur, and M. M. Hassan. Spam message detection using danger theory and krill herd optimization. *Computer Networks*, 199 :108453, 2021.

K. B. Stevens, M. Gilbert, and D. U. Pfeiffer. Modeling habitat suitability for occurrence of highly pathogenic avian influenza virus h5n1 in domestic poultry in asia : a spatial multicriteria decision analysis approach. *Spatial and spatio-temporal epidemiology*, 4 :1–14, 2013.

S. Valentin, E. Arsevska, S. Falala, J. De Goër, R. Lancelot, A. Mercier, J. Rabatel, and M. Roche. Padi-web : A multilingual event-based surveillance system for monitoring animal infectious diseases. *Computers and Electronics in Agriculture*, 169 :105163, 2020.

S. Yousefinaghani, R. A. Dara, Z. Poljak, and S. Sharif. A decision support framework for prediction of avian influenza. *Scientific Reports*, 10(1) :1–14, 2020.

W. Zhou and Y. Liang. A new version of the deterministic dendritic cell algorithm based on numerical differential and immune response. *Applied Soft Computing*, 102 :107055, 2021.

W. Zhou, Y. Liang, Z. Ming, and H. Dong. Earthquake prediction model based on danger theory in artificial immunity. *Neural Network World*, 30(4) :231, 2020.