Integrating Epidemiological and Environmental data for Enhanced Event-Based Surveillance Systems

Bahdja Boudoua

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'UNIVERSITÉ DE MONTPELLIER

En Informatique

École doctorale Information, Structures, Systèmes (I2S)

Unité de recherche TETIS, INRAE

Integrating Epidemiological and Environmental Data for Enhanced Event-Based Surveillance Systems

Présentée par Boudoua El Bahdja

Sous la direction de Maguelonne Teisseire, Annelise Tran et Mathieu Roche

Devant le jury composé de

Emmanuelle Gilot-Fromont, Professeure, VetAgro Sup, France	Rapportrice
Sabine Loudcher, Professeure, Université Lyon 2, France	Rapportrice
Benoit Durand, Chercheur, Anses, France	Examinateur
Isabelle Mougenot, Professeure, Université de Montpellier, France	Examinatrice
Maguelonne Teisseire, Directrice de Recherche INRAE, France	Directrice de thèse
Annelise Tran, Chercheuse CIRAD, France	Co-Directrice de thèse
Mathieu Roche, Chercheur CIRAD, France	Invité, Encadrant

UNIVERSITÉ DE MONTPELLIER

PUBLICATIONS, DATA AND CODE

Journals

Bahdja Boudoua, Mathieu Roche, Maguelonne Teisseire, Annelise Tran, *EpiDCA: Adaptation and implementation of a danger theory algorithm for event-based epidemiological surveillance, a case study of avian influenza*, Computers and Electronics in Agriculture, Volume 229, 2024, [Q1, IF 7.7]
https://doi.org/10.1016/j.compag.2024.109693

Valentin S, **Boudoua B**, Sewalk K, Roche M, Lancelot R, Arsevska E, *Analysis of the flow of information between sources used by event-based surveillance systems. Case study of Avian influenza using HealthMap and PADI-Web*, PlosOne 18, no. 9, 2023 [Q1, IF 2.9]
https://doi.org/10.1371/journal.pone.0285341

Arınık N, Van Bortel W, **Boudoua B**, Busani L, Decoupes R, Interdonato R, Kafando R, van Kleef E, Roche M, Syed MA, Teisseire M, An annotated dataset for event-based surveillance of antimicrobial resistance. *Data in Brief*. Elsevier, Volume 46, 2023 [Q2, IF 1.2]
https://doi.org/10.1016/j.dib.2022.108870

Conference Proceedings

Boudoua B, Roche M, Teisseire M, Tran A. *International Conference of Spatial Epidemiology, Geostatistics and GIS applied to animal health, public health and food safety, GeoVet, Teramo, Italy, 2023.* How to combine spatio-temporal information and Danger theory for animal disease surveillance?"
https://www.veterinariaitaliana.izs.it/index.php/GEOVET23/article/view/3301

Boudoua B, Roche M, Teisseire M, Tran A. *EGC (Extraction Gestion des connaissance), Lyon, 2022.* EpiDCA : Théorie du danger, veille sanitaire et facteurs de risques. vol. RNTI-E-39, pp.633-634 https://editions-rnti.fr/?inprocid=1002872

Boudoua B, Hautefeuille C., Arsevska E, Valentin S, Understanding Outbreak Data Dissemination In Event Based Surveillance Systems. Application On Avian Influenza Using PADI-web. *International Journal of Infectious Diseases*. 116, suppl.: S99. International Journal of Infectious Diseases (IMED 2021)
https://doi.org/10.1016/j.ijid.2021.12.234

Production of datasets

Boudoua B, Richard, M, Roche M, Teisseire M, Tran A. 2023, *Annotated datasets from PADI-web for event-based surveillance of Avian Influenza, African Swine Fever, and West-Nile Virus Disease*
https://doi.org/10.57745/99SNOZ,RechercheDataGouv,V1

Boudoua, B, Tran, A. 2023, *Suitability map for Avian influenza, Asia*
https://doi.org/10.18167/DVN1/FYWDOJ,CIRADDataverse,V1

Boudoua, B, Richard, M, Roche, M, Teisseire, M, Tran, A. 2023, *Annotated datasets from PADI-web for event-based surveillance of Avian Influenza, African Swine Fever, and West-Nile Virus Disease*
https://doi.org/10.57745/99SNOZ,RechercheDataGouv,V1

Nejat, A, Van Bortel W, **Boudoua B**, Busani L, Decoupes R, Interdonato R, Van Kleef E, Kafando R, Roche M, Syed MA, Teisseire, M, 2022, *MOOD - News AMR dataset - Hackathon 2022*
https://doi.org/10.57745/MPNSPH,RechercheDataGouv,V4

Valentin S, **Boudoua B**, Sewalk K, Arsevska E, 2022 *Dissemination of information in event-based surveillance, a case study of Avian Influenza - Dataset*
https://zenodo.org/record/7828530

Code

EpiDCA method in R language
https://github.com/BBahdja/EpiDCA.git
Code to extract coordinates (latitude, longitude) from text (place name)
https://github.com/BBahdja/Location_Extraction.git

RÉSUMÉ

Les systèmes de surveillance basée sur les événements (SBE) tels que HealthMap, ProMED et PADI-web sont utilisés quotidiennement afin de détecter des événements épidémiologiques signalés dans les médias en ligne (articles). Une fois les articles collectés, ces systèmes s'appuient sur des algorithmes de classification supervisée et/ou une modération humaine pour classer les articles selon leur pertinence. L'application de telles méthodes peut être difficile, car les jeux de données épidémiologiques ne sont pas équilibrés. D'autre part, l'annotation d'articles, qui sert à l'apprentissage des méthodes supervisées, est coûteuse et chronophage. De plus, les facteurs de risque liés à l'apparition et transmission des maladies (facteurs de risque environnementaux et épidémiologiques) ne se trouvent pas toujours dans les données textuelles et ne sont donc pas pris en compte par les systèmes de SBE.

Dans ce contexte, nous proposons une approche non-supervisée qui s'appuie sur les informations spatio-temporelles des événements épidémiologiques détectés, pour classer les articles en tenant compte des facteurs environnementaux par le biais de cartes de risques. Cette méthode, appelée EpiDCA, est une adaptation de l'algorithme des cellules dendritiques (DCA), inspirée par la théorie du danger. EpiDCA se caractérise par des paramètres définis par des experts, ce qui le rend applicable à différentes maladies et contextes environnementaux. La méthode proposée a été testée sur un premier jeu de données relatif à l'influenza aviaire en Asie entre 2018 et 2019, ainsi qu'une carte de risque produite pour la même région. Pour l'évaluer, nous avons calculé la précision, le rappel et le F-score. EpiDCA a obtenu une très bonne performance avec un F-score de 0,82 pour un jeu de données déséquilibré et de 0,90 pour un ensemble de données équilibré. Les résultats ont également confirmé que la prise en compte des facteurs de risque des maladies est une bonne approche pour la classification des événements. EpiDCA a ensuite été comparé aux méthodes d'apprentissage supervisé et s'est avéré compétitif.

Après cette application initiale, l'objectif était d'évaluer la robustesse et la généricité de la méthode dans différents contextes géographiques et à travers divers systèmes épidémiologiques, notamment une maladie animale transfrontalière (la peste porcine africaine) et une maladie zoonotique vectorielle (la fièvre du Nil occidental) en Europe. Nous avons construit un jeu de données original à partir des articles détectés par PADI-web. Nous avons également développé une méthode d'annotation pour labelliser les articles. Nous avons ensuite proposé une extension de la méthode qui permet d'intégrer des covariables supplémentaires pour l'améliorer en termes de réactivité et de précision. Les perspectives avec EpiDCA incluent la réduction du nombre de paramètres et l'application du modèle à d'autres contextes de surveillance qui s'appuient sur les mêmes types de sources, tels que les maladies végétales et la sécurité alimentaire.

Mots-clés : Surveillance basée sur les événements, Théorie du danger, Algorithme des cellules dendritiques, Influenza aviaire, Peste porcine africaine, Fièvre du Nil occidental.

ABSTRACT

Event Based Surveillance (EBS) systems such as HealthMap, Promed and PADI-web are used daily to timely detect outbreak events reported in web articles. Once the articles are collected, these systems rely on human moderation and supervised classification algorithms to classify articles according to their relevance. Applying such methods can be challenging, as epidemiological datasets have an imbalanced class distribution, and because the annotation task, which is critical to the success of these models, can be expensive and time consuming. Another important limitation of EBS systems is that the drivers of disease transmission (e.g. disease characteristics, environmental and epidemiological risk factors) are not always found in textual data and are therefore not taken into account by EBS systems.

In this context, we propose an unsupervised approach that relies on the spatio-temporal information of the reported epidemiological events, to classify articles while taking into account the environmental factors associated with disease onset through risk mapping. This method, called EpiDCA, is an adaptation of the Dendritic Cells Algorithm (DCA), inspired by the danger theory. EpiDCA is characterized by expert-defined parameters, making it applicable to different diseases and environmental contexts. The proposed method was first tested and evaluated using a dataset related to avian influenza (AI) in Asia between 2018 and 2019, and a suitability map for AI produced for the same area. To measure the accuracy of the model, we calculated the precision, recall and F-score. EpiDCA achieved a very good performance with an F-score of 0.82 and 0.90 for an imbalanced and a balanced dataset respectively. The results confirmed that considering disease risk factors is a good approach in event classification. EpiDCA was then compared with state-of-the-art supervised machine learning methods and appeared to be competitive.

After this initial application, we aimed to evaluate the robustness and genericity of the method in different geographical contexts and across various epidemiological systems, specifically; a transboundary animal disease (African Swine Fever) and a vector-borne zoonotic disease (West-Nile Virus Disease) in Europe. For this purpose, we constructed an original dataset from articles detected by PADI-web. We also developed a method and guidelines to annotate the articles. The consistent results confirmed the robustness of EpiDCA. Then we extended the method by integrating additional covariates to further enhance its reactivity and accuracy. Future perspectives with EpiDCA include the reduction of the number of parameters and the application of the model to other surveillance contexts that rely on the same sources, such as plant disease surveillance, and food security surveillance.

Keywords: Event-based surveillance, Danger theory, Dendritic Cells Algorithm, Avian Influenza, African Swine Fever, West-Nile Disease.

ACKNOWLEDGEMENT

Écrire les remerciements s'est avéré plus difficile que je ne le pensais, car les mots me manquent pour exprimer ma gratitude à toutes les personnes avec qui j'ai pu échanger et partager du temps depuis mon arrivée à Montpellier. Bien que toutes ne soient pas mentionnées dans ces remerciements, chacune a contribué à faire de cette aventure une expérience enrichissante pour moi aussi bien sur le plan scientifique que personnel.

Au cours de ce parcours parfois laborieux qu'est la thèse, j'ai eu la chance et l'honneur d'être entourée par une équipe d'encadrants exceptionnels. Maguelonne, Annelise et Mathieu, ce fut un véritable plaisir de collaborer avec vous pendant ces trois années. Merci de m'avoir guidée avec exigence et bienveillance, de m'avoir fait confiance et de m'avoir transmis votre passion de la recherche et de vos domaines respectifs. Vous resterez des modèles pour moi durant toute ma carrière et j'espère continuer à échanger avec vous.

Je voudrais exprimer ma profonde gratitude à ma directrice de thèse, Maguelonne Teisseire, pour m'avoir guidée et transmis son expertise, toujours dans la bonne humeur, pour sa disponibilité, pour avoir cru en moi, et pour m'avoir secouée quand il le fallait. Mes sincères remerciements à ma co-directrice de thèse Annelise Tran, pour m'avoir transmis son savoir avec patience et bienveillance, pour sa présence et pour son soutien. Mes sincères remerciements à mon encadrant, Mathieu Roche, pour son implication dans mon travail de thèse à toute heure, pour sa disponibilité, et pour sa patience lorsqu'il a parfois fallu me répéter la même chose plusieurs fois. Je vous remercie également pour tous les moments que nous avons partagés durant nos missions. Cela a été un plaisir de voyager et de découvrir d'autres pays, villes et spécialités culinaires avec vous.

Je tiens à remercier le projet MOOD (Monitoring Outbreaks for Disease surveillance in a data science context) pour le soutien financier, qui a été essentiel dans la réalisation de ma thèse et grâce auquel j'ai pu mener à bien mes recherches et bénéficier de ressources précieuses qui ont enrichies mon expérience académique.

Je suis également très reconnaissante envers l'INRAE (UMR TETIS) et la Maison de la Télédétection pour les ressources, l'infrastructure et l'environnement mis à disposition, qui ont facilité le bon déroulement de mon travail et l'ont rendu agréable. Je voudrais également remercier chaleureusement tous mes collègues et amis pour leur soutien indéfectible, ainsi que pour les conversations stimulantes et inspirantes que nous avons partagées tout au long de ce parcours. Merci à Alexandre, Babacar, Claire, Clovis, Cyrille, Emmanuel, Karine, Larisa, Maksim, Pauline, Remy, Renaud, Roberto, Romain et Thibault. Merci à Sarah et Elena, qui ont été les premières à m'initier au monde fascinant de la surveillance basée sur les évènements. Merci à Carlène pour sa disponibilité, son aide et son expertise. Merci à Manon pour son travail et son enthousiasme durant son stage.

Je remercie du fond du cœur ma famille, mon papa qui m'a toujours soutenue, qui m'a appris à faire face à n'importe quelle situation avec patience, résilience et humour, je sais que tu veilles toujours sur nous et que tu nous guides de là où tu es. Merci maman, d'être notre

pilier, d'avoir les mots justes et les bons conseils en toutes circonstances. Merci également à ma sœur Sarah et à mon frère Abdelhakim (Kimo) pour leur soutien inconditionnel.
Pour finir, merci à la préfecture pour avoir alimenté ma ténacité et ma créativité face aux problèmes administratifs. Merci à mon voisin Philippe et à la police de Montpellier pour avoir empêché le vol de mon ordinateur, sans votre intervention cette thèse aurait certainement pris plus de temps que prévu.
Puisque j'écris ces quelques lignes après ma soutenance, je remercie encore une fois toutes les personnes présentes ce jour-là, dans la salle ou en visio. Merci d'avoir fait du 14 octobre 2024 un souvenir inoubliable pour moi.

List Of Abbreviations

ADNS	Animal Disease Notification System
AGs	Antigens
AI	Avian Influenza
ASF	African Swine Fever
BERT	Bidirectional Encoder Representations from Transformers
COVID-19	Coronavirus Disease 2019
CSM	Cumulative Output Signal
DANIEL	Data Analysis for Information Extraction in Any Language
DCA	Dendritic Cells Algorithm
DCs	Dendritic Cells
Ds	Danger signals
EBS	Event-based Surveillance
ECDC	European Centre for Disease Prevention and Control
ECMWF	European Centre for Medium-Range Weather Forecasts
EMMA	European Media Monitor Alerts
ESA	Épidémiosurveillance en Santé Animale
EWS	Early Warning System
FAO	Food and Agriculture Organization
GEE	Google Earth Engine
GIS	Geographical Information System
GPHIN	Global Public Health Intelligence Network
GRITS	Global Rapid Identification System for Infectious Diseases in Textual Data Sources
HPAI	Highly Pathogenic Avian Influenza
IBS	Indicator-Based surveillance
IE	Information extraction
ISID	International Society for Infectious Diseases
JE	Japanese encephalitis
K-nn	K-nearest neighbors
LPAI	Low Pathogenic Avian Influenza
MedISys	Medical Information System
MMWR	Morbidity and Mortality Weekly Report
MT	Migration Threshold
NLP	Natural Language Processing
OIE	World Organization for Animal Health
PADI-web	Platform for Automated extraction of animal Disease Information from the web
ProMED	Program for Monitoring Emerging Diseases
RF	Random Forest
SARS	Severe Acute Respiratory Syndrome
Ss	Safe signals
SVM	Support Vector Machine
VSI	Veille Sanitaire Internationale
WAHIS	World Animal Health Information Database Interface
WHO	World Health Organization
WND	West Nile Disease
WOAH	World Organisation for Animal Health

CONTENTS

List of Figures

LIST OF TABLES

Introduction

This chapter provides an understanding of Event-Based Surveillance systems within the context of Epidemic Intelligence (EI), establishing the global framework for this thesis. It highlights the research objectives and summarizes the contributions associated with these objectives.

Emerging diseases represent a growing risk for both public health and veterinary health [36]. To mitigate the risk of outbreaks, many countries have adopted an Epidemic Intelligence (EI) strategy that integrates two components [123]: i) indicator-based surveillance (IBS) relying on official sources such as the World Health Organisation (WHO), the World Organisation for Animal Health (WOAH), or the Food and Agriculture Organisation (FAO), and ii) Event-based surveillance (EBS) relying on unofficial sources (online media, social networks, etc.).

IBS systems produce structured and reliable data, offering an extensive range of information regarding confirmed epidemiological events. However, these systems can be limited in their ability to detect early events due to the delays inherent in the pipeline process, which includes disease observation, laboratory confirmation, administrative processing, and final reporting [146]. EBS, on the other hand, is the organized process of detecting and reporting information, represented as events, to public health authorities by rapidly capturing data from various unstructured sources [11]. This system enables authorities to be better prepared for endemic and pandemic disease outbreaks by serving as a crucial component of an effective early warning system [11, 122]. Together, IBS and EBS systems complement each other by addressing different needs in disease surveillance. While IBS systems provide detailed and structured data about confirmed cases, EBS systems offer rapid detection and initial alerts.

Since the early 2000s, several automatised EBS tools that are now called EBS systems, have been developed to collect and analyze a continuous stream of unstructured textual data, such as news articles and reports, to extract timely and relevant information about outbreaks and events [57, 161, 29]. The final output of all the EBS systems is a set of articles classified according to their relevance to the epidemiological topic, with epidemiological data extracted from these articles. Various studies have since assessed their performances, and limitations. Their capacity to detect relevant health information has been recognised [12]. For example, during the SARS outbreak (Severe Acute Respiratory Syndrome) in 2003 [115], the H1N1 outbreak (a strain of influenza) in 2009 [88], and more recently the COVID-19 pandemic (Coronavirus Disease 2019) [20].

EBS systems can be classified based on the diseases they cover (public health, animal health, plant health or one health), the languages they support, their geographical focus, or, more importantly, the type of moderation they employ. Moderation can vary: it can be manual, meaning human-moderated, as in the case of ProMED (Program for Monitoring Emerging Diseases) managed by the International Society for Infectious Diseases (ISID) [171], semi-automated (a hybrid of manual and automated processes) like the Global Public Health Intelligence Network (GPHIN) and HealthMap [57], or fully automated, as it is for European Commission Medical Information System (MedISys) [133], and the Platform for Automated Extraction of Disease Information (PADI-web) [161]. Each method has its strengths and limitations. ProMED outputs present a very low level of false positive detection but it is limited by resource constraints (availability of experts), and this expert validation is time-consuming

[29]. Semi-automated systems, combining automated data collection and classification with expert moderation, offer improved timeliness but still facing similar limitations as fully manual systems. Unlike moderated systems, fully automated systems process data more quickly and are more cost-efficient, thanks to machine learning classifiers. However, they also face limitations, such as dealing with noisy data and filtering out false positives.

Motivation

Given the overwhelming amount of data available on the web, one of the primary challenges to establishing and sustaining an EBS system is designing a system that can detect a sufficient number of relevant health events while ensuring it is not overloaded [11]. EBS systems usually classify collected articles as relevant or irrelevant by relying on human moderation or by implementing classification algorithms. These systems use annotated data to improve their classification in terms of accuracy and thus swiftly detect outbreak events. Consequently, the performance of these algorithms is highly dependent on the quality of the dataset used to train them [114]. Indeed, epidemiological text classification can be challenging for various reasons [156]. First, epidemiology related texts can be ambiguous, as sometimes the disease is mentioned but none outbreak is reported. Instead, they might present general information related to the disease, or draw the disease history in a given area [157]. Second, a single article can report one event (outbreak) and simultaneously report other events or other types of epidemiological information in other areas. In this case, different types of information and multiple locations are found within the same text.
Notably, disease-related characteristics and environmental drivers are not always found in textual documents and are therefore not taken into account in the classification. However, disease-related characteristics and environmental drivers can significantly influence the way information is processed and reported. For example, West-Nile virus Disease (WND) outbreaks are more likely to occur in summer when vectors are active, making events reported during that period more likely to be relevant [66]. To the best of our knowledge, no research study has attempted to address the classification limitations in event-based surveillance systems by combining epidemiological and environmental data.
This thesis is interdisciplinary and enables the combination of data-driven approaches and model-based methods that integrate expert knowledge.

This thesis is mainly funded by the '**MOnitoring Outbreaks for Disease surveillance H2020 in a data science context (MOOD[1])**' project. The MOOD project aims at taking advantage of data mining, analysis and visualization of health, environmental and other data to **enhance the utility of EBS**. Ultimately, MOOD is supporting the work of European and global public and veterinary health agencies and surveillance practitioners by providing

[1] https://mood-h2020.eu/

existing monitoring platforms with novel features, and methodological and practical support adapted to their needs.

Objectives

Our main objective is to address a key limitations of EBS systems such as managing the overwhelming volume of collected articles and addressing the lack of explainability for detected events. We propose an unsupervised model that is robust and generic (independent of a specific disease or host) and that integrates epidemiological and environmental data, placing detected epidemiological events in their environmental context. This approach not only enhances the classification of EBS systems but also provides valuable explanatory insights and incorporates expert epidemiological knowledge. Our model is inspired by the Dendritic Cells Algorithm (DCA) [67], which is based on the danger theory [105]. This choice is justified by DCA's advantages in real-time applications: it operates in an unsupervised manner without requiring training periods [172] and has proven effective in reducing false positives [111]. To the best of our knowledge, applying DCA within the context of EBS systems is novel and has not been explored before.

In this context, we want to address the following research questions:

- How can the DCA be applied to EBS systems, and how can its inherent limitations be addressed?
- What specific types of epidemiological and environmental information should be used as inputs for the method?
- How can we evaluate the robustness and genericity of the method through case studies?

Contributions

In this thesis, we present four key contributions to address our objectives:

- Development of EpiDCA, an unsupervised method based on the Dendritic Cells algorithm (DCA) to combine epidemiological and environmental data in EBS systems.
- Establishment of an annotation method and production of an original dataset for three different case studies to evaluate the proposed method.
- Evaluation of the method's robustness and genericity using the produced dataset.

- Introduction of an extension of EpiDCA designed to integrate real-time environmental data into the model.

This thesis is organized into two main parts. **Part I: State of the Art** begins with Chapter 1, which reviews the state of the art in various surveillance approaches, including IBS and EBS. Chapter 2 details the characteristics and epidemiological context of the diseases used as case studies. Chapter 3 discusses the inspiration behind the proposed method.
Part II: Contributions starts with Chapter 4, which introduces EpiDCA, the method developed, and its application to an initial case study. Chapter 5 details the annotation method applied to create an original dataset. This dataset is used in Chapter 6 to evaluate the method. Finally, Chapter 7 explores an extension of the proposed approach. Chapter 8 concludes the thesis by summarizing the key findings, drawing conclusions, and discussing future perspectives.

Part I

State of the art

CHAPTER 1

DATA-DRIVEN EPIDEMIOLOGICAL SURVEILLANCE

In this chapter, we aim to set the context of our research and provide a comprehensive overview of current surveillance approaches. We begin by presenting Epidemic Intelligence (EI) as a framework for disease surveillance, introducing Indicator-Based Surveillance along with its main characteristics and limitations. Next, we explore event-based surveillance (EBS), which is the primary focus of our research, discussing its associated systems and highlighting current challenges. Finally, we discuss potential strategies to address these challenges.

1.1 Introduction

Epidemiological surveillance relies on an intersection of various overlapping strategies. They differ in terms of the nature of the data collected and the types of sources used, but they are complementary and they converge toward the same ultimate goal: the timely detection of events representing a threat to human and animal health.
EI concept, as it is used today, was developed in the early 2000s. The French Institut de Veille Sanitaire (Institute of Health Surveillance) and the European Center for Disease Prevention and Control (ECDC) proposed an EI framework to enhance disease surveillance in Europe in 2006 [156]. Eight years later, the World Health Organisation (WHO) published a comprehensive guide providing key definitions and detailing the implementation of early warning activities [122]. EI can be defined as a formalized surveillance process that encompasses all activities related to the early identification of potential health hazards that may represent a risk to health, and their verification, assessment and investigation, EI relies on two main and complementary components: "Indicator-based surveillance" which refers to structured data collected through routine surveillance systems and "Event-based surveillance" which refers to unstructured data on potential and non-verified disease outbreaks (i.e. events) gathered from sources of any nature.
In this chapter, we will discuss Indicator-Based Surveillance (IBS) and then focus on Event-Based Surveillance (EBS) systems. This will set the context of our work, including the limitations we aim to address.

1.2 Indicators-based surveillance

IBS involves the systematic collection, monitoring, analysis, and interpretation of structured data, such as case numbers, prevalence rates [1], and mortality rates[2]. This data originates from official sources at various levels, with local public health authorities reporting to national agencies, which in turn contribute data to international organizations that centralize and analyze information on a global scale [50, 156].
On a global scale, three main health organizations play a crucial role in coordinating surveillance and sharing information:

- **World Organisation for Animal Health (WOAH)**: Established in 1924 to combat infectious animal diseases [26].

[1] Prevalence rate is the proportion of a population that has a specific disease or condition at a particular time or over a specified period.

[2] Mortality rate is defined as the number of deaths in a given population during a specific time period divided by the total population.

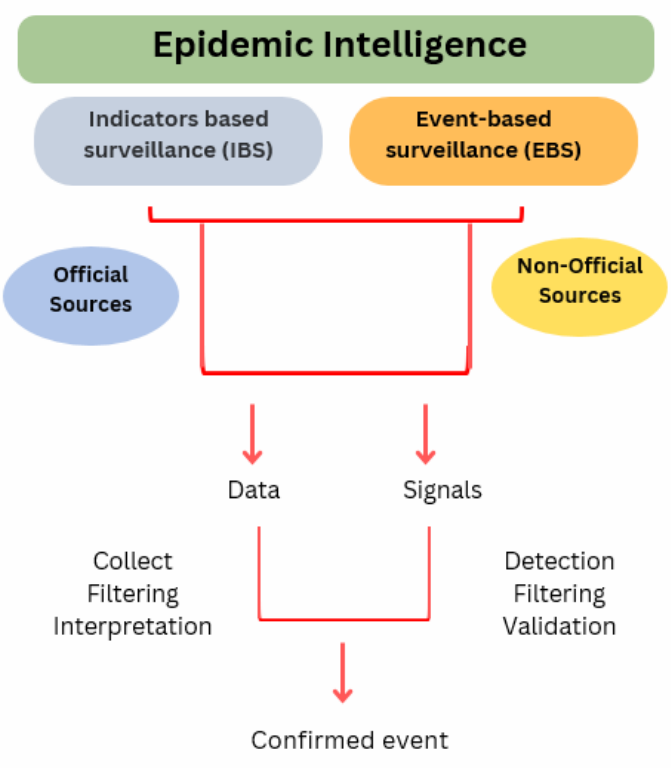


Figure 1.1: Epidemic Intelligence (EI) Framework.

- **World Health Organization (WHO)**: Founded in 1948 to monitor human transmissible diseases, including zoonoses.
- **Food and Agriculture Organization (FAO)**: Created in 1945 to improve agricultural productivity in developing countries, notably through the enhancement of veterinary services [107].

1.2.1 Indicator-based surveillance systems

After the 1990s, information technology, such as web access, online databases, and the development of geographic information systems, significantly evolved, pushing international organizations to implement more transparent disease reporting systems via appropriate web platforms [6]. In 1994, the FAO launched an emergency prevention and rapid response system for transboundary diseases (EMPRES), initially focusing on swine plagues and avian influenza. Since 2004, EMPRES-i, the FAO's web platform open to the community, has allowed the visualization of epidemiological data on more than 34 transboundary diseases [164]. In 1996, the OIE implemented a secure online disease reporting system. This system was modernized in 2006 with a web interface open to the community (WAHIS: World Animal Health Information System). This interface provides access to epidemiological information on more than 100 infectious diseases of terrestrial and aquatic animals, reported

in over 180 countries [99]. At the European level, two systems are responsible for collecting, centralizing, and sharing health data on infectious diseases. The Animal Disease Notification System (ADNS) [65], created in 1982, centralizes and analyzes health data on 45 exotic animal diseases that may emerge in Europe in order to alert European countries in case of an introduction risk. Similarly, the online notification system TESSy, created in 2004, allows the sharing of health information on 52 human infectious diseases [2]. IBS systems offer distinct advantages by presenting officially confirmed (validated) information through maps and graphs, facilitating visual representation of disease outbreaks, as shown in Figure 1.2. They communicate findings via reports and provide users with the ability to download data in structured tabular formats containing detailed epidemiological information, such as the location, date of observation and confirmation, host, number of host and more. The characteristics of these main IBS systems are detailed in Table 1.1, providing comparison of their creation dates, geographic coverage, languages, targeted diseases, information sources, and access levels 1.1.

IBS system	WAHIS	ADNS	EMPRES-i	TESSy
Year of Creation	1996	1998	2004	2004
Geographic Coverage	Global	European	Global	European
Number of Languages	3	1	1	1
Targeted Disease[1]	A	A	A, H	H
Sources[2]	O	O, N	O	O
Access	Public	Restricted	Public	Restricted and Public
References	[99]	[65]	[164]	[2]

[1] A = Animal, H = Human.
[2] O = Official N = Non-official.

Table 1.1: Characteristics of the main IBS systems adapted from [4].

These features make IBS systems invaluable as a reference or gold-standard for evaluating EBS systems (that will be presented in Section 1.3). They are essential for assessing detection accuracy (i.e. identification of confirmed events) and for measuring reactivity (i.e. the time difference between detection by EBS systems and confirmation by IBS systems).

1.2.2 Limitations

While IBS systems and official sources remain the cornerstone of disease surveillance, like all systems, they have their limitations. First, before reporting to an international health authority, they follow a specific pipeline: disease observation, laboratory confirmation, administrative processing, and final reporting, often resulting in an unavoidable delay [146]. When it comes to the timely detection of outbreaks and important public health events, IBS systems often fail, as presented in a retrospective study by [78] on H1N1 outbreaks in 2009, and

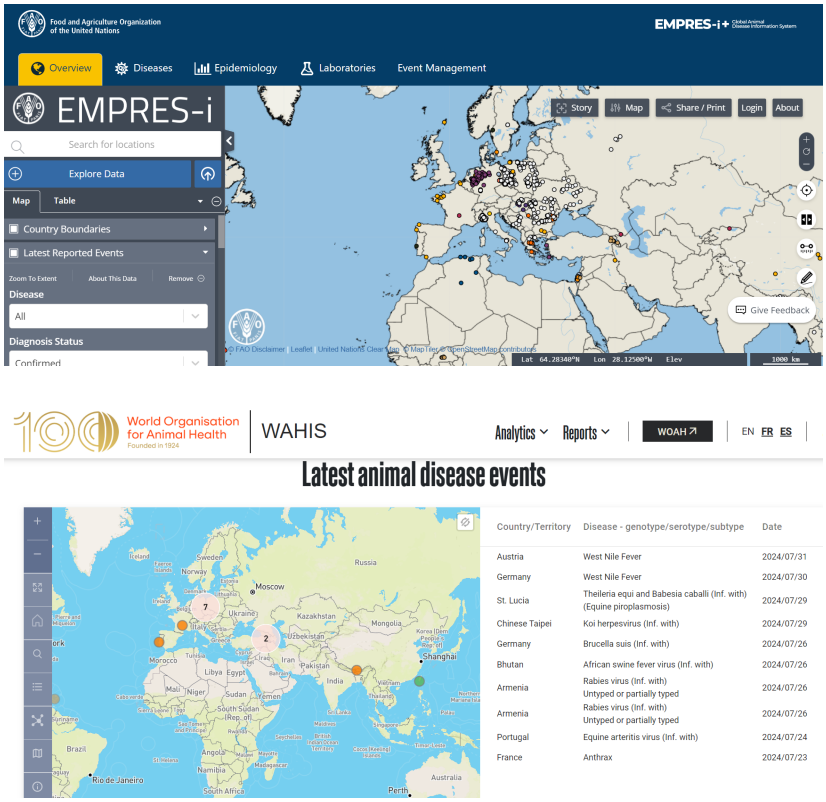


Figure 1.2: Events display on EMPRES-i and WAHIS platforms.

Zika outbreaks in 2015. Second, in resource-limited settings, classical IBS approaches can be limited by available diagnostic capacity and surveillance architecture [15]. Furthermore, the systems are not suited to the detection of rare but high-impact outbreaks or emerging and unknown diseases several examples has highlighted this limitation, such as Avian Influenza and COVID-19 outbreaks [121].

1.3 Event-based surveillance

Event-based surveillance (EBS) have been developed to address the limitations of indicator-based surveillance (IBS) [13] and complement it by relying on non-official sources, such as web articles, social media, and other digital platforms, to detect events [59, 89]. An event in this context is defined as any occurrence that may pose a threat to public health and requires urgent attention [11]. EBS involves the daily collection, monitoring, assessment, and interpretation of potential reported outbreaks from a variety of sources. Systems like ProMED [29], HealthMap [57], MedISys, GPHIN, Argus, BioCaster and PADI-web [161] are designed to detect unusual health events reported in web articles and extract relevant health information. These systems enable near real-time detection of infectious disease outbreaks by identifying and analyzing relevant articles on a daily basis [59].

1.3.1 Event-based surveillance systems

ProMED, established in 1994, is one of the first event-EBS systems implemented by the International Society for Infectious Diseases (ISID). It enables health practitioners and the public to report potential infectious disease outbreaks. Reports, both formal and informal, are reviewed and commented on by subject matter experts before being posted to the global network. ProMED's reports, which focus on emerging and re-emerging outbreaks as defined by the WHO, are accessible to over 90,000 subscribers and followers worldwide [38].
GPHIN, short for the Global Public Health Intelligence Network, was developed in 1997 through a partnership between the Canadian government and the WHO. It functions as a multilingual EBS system, its primary role is to gather and disseminate relevant information on disease outbreaks and other public health events by monitoring global media sources including news wires and websites [106].
Argus and MedISys both lunched in 2004 represent significant advancements in EBS. Developed by the Center for Infectious Disease Research at Georgetown University, Argus focuses on identifying potential health threats within the United States. Meanwhile, MedISys, initiated by the Joint Research Center at the request of the European Commission, serves as an automatic news aggregator with extensive global coverage across more than five thousand topics, spanning animal health, public health issues, and threats related to chemical, nuclear, and bio terrorist attacks [133, 117].
HealthMap, developed by Harvard University in 2006, collects and integrates outbreak data from a variety of sources, including news media (e.g., Google News), expert-curated accounts (e.g., ProMED), and validated official alerts. Through the use of text processing algorithms, the system classifies alerts by location and disease and then overlays them on an interactive geographic map [57]. The same year the university of Tokyo launched BioCaster, a non-governmental public health surveillance system known for its open ontology-centered approach, focusing on Asia-Pacific languages [35].
In France, since 2016, the International health monitoring (Veille Sanitaire Internationale, VSI) of the Animal Health Epidemiological Surveillance Platform (Épidémiosurveillance en Santé Animale, ESA) has been using PADI-web (Platform for Automated extraction of Disease Information from the Web) to complement its event-based surveillance component. PADI-web ensures the detection, verification, and communication of infectious disease signals. Unlike the previously mentioned health surveillance systems, PADI-web has been developed initially for animal health surveillance [6], and recently for plant disease surveillance [132].

While this is not an exhaustive review of EBS systems, these ones are often cited as examples or used in research studies due to their extensive geographical and disease coverage, operational platforms, and accessible data. Other EBS systems exist and are currently under development. For example, recent advancements include the Global Rapid Identifi-

cation of Threats System for Infectious Diseases in Textual Data Sources (GRITS) [79] and the Data Analysis for Information Extraction in Any Language (DANIEL). GRITS enhances epidemic surveillance by automatically analyzing epidemiological texts to extract critical information about disease outbreaks, such as the likely disease, dates, and affected countries, with the innovative option of suggesting potentially associated infectious diseases. Similarly, DANIEL's novel approach allows the system to process multiple languages without the need for translations. The benefit of this method is to increase coverage across a variety of languages, including low-resourced languages, rather than focusing on optimizing results for a specific language [134].

Typically, all EBS systems revolve around four key steps: data collection, classification, information extraction, and communication. The global framework of the EBS system is illustrated in Figure 1.3. We describe the steps involved: data collection, classification, information extraction, and communication, using examples. However, it's important to note that these steps are not always distinct. For instance, extraction and classification might occur in different orders, and extraction may not be used in all EBS systems. For simplicity and clarity, we present them in this specific sequence to clearly illustrate each step

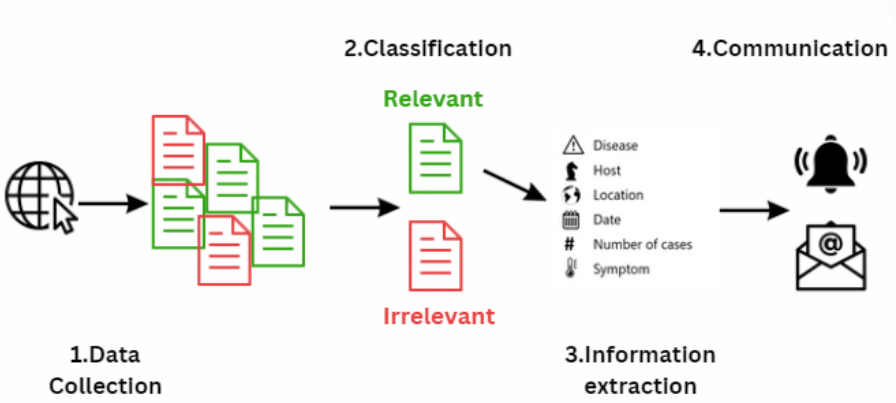


Figure 1.3: Key steps of event-based surveillance systems.

Data collection

This step involves the daily detection and collection of epidemiological events, with each system using a variety of methods and sources, which may overlap. ProMED for example, identifies potential infectious disease outbreaks through reports submitted by both formal sources, such as health practitioners and public health professionals, and informal sources, including concerned members of the public. It operates in multiple languages including French, Spanish, Portuguese, Russian and Arabic [171, 29].

HealthMap collects and integrates data from informal sources, including news articles, social platforms aggregated through Google News, Baidu, SOSO and formal souces like expert-curated accounts from ProMED [57, 25]. MedISys retrieves news articles from specialised

official and unofficial medical sites, general news media and selected blogs, it operates in over 50 languages, making it one of the most expansive systems of its kind to date [133]. Argus [116] collects information in 40 native languages from media sources, including printed newspapers, electronic media, Internet-based newsletters and blogs, as well as from official sources the WHO and WOAH [100]. BioCaster [35] collects information from EurekAlert!, European Media Monitor Alerts (EMMA), Google, the CDC's Morbidity and Mortality Weekly Report (MMWR), MeltWater, WOAH, ProMED, Reuters, WHO and Vetsweb. It scans for articles in Arabic, Chinese, English, French, Japanese, Korean, Portuguese, Russian, Spanish, Thai and Vietnamese. The system gives a special priority to languages of the Asia-Pacific region.
PADI-web retrieves web articles from Google News through two types of customized RSS feeds: Disease-based RSS feeds consist of disease names, while symptom-based RSS feeds include clinical signs and hosts. It operates in 16 languages [156]. Table 1.2 presents the main characteristics of the EBS cited in this section, including disease and geographical coverage, targeted languages, and sources used.

EBS system	ProMED	GPHIN	Argus	MedISys	HealthMap	BioCaster	PADI-web	GRITS	DANIEL
Year of Creation	1994	1997	2004	2004	2006	2006	2016	2012	2015
Geographic Coverage	Global	Global	Global	Global	Global	Global[1]	Global	Global	Global
Number of Languages [2]	5	9	40	50	7	13	M	M	M
Targeted Disease[3]	H, A, P	H, A, P, E	H, A, P, E	H, A, P, E	H, A, P, E	H, A, P, E	A, P	H, A	H, A
Classification[4]	H	SA	A	H	SA	SA	A	A	A
Sources[5]	O, N	O, N	O, N	N	N	O, N	N	N	N
Type of Sources[6]	W, U	W	W	W	W, U, S	W, U, S	W	W	W
References	[171, 29]	[115]	[117]	[133]	[57]	[35]	[161]	[79]	[93]

[1] BioCaster focuses on Asia-Pacific languages and health hazards.
[2] M = Multilingual [3] H = Human, A = Animal, P = Plant, E = Environmental.
[4] H = Human intervention (experts), SA = Semi-automated, A = Automated.
[5] O = Official, N = Non-Official.
[6] W = Web, S = Social media, U = Users.

Table 1.2: Characteristics of the main EBS systems, adapted from [4].

Classification

The classification step is a crucial step in EBS. Given the overwhelming amount of data available on the web, one of the primary challenges to establishing and sustaining an EBS system is designing a system that can detect a sufficient number of relevant health events while ensuring it is not overloaded [11]. The documents classification can be manual i.e., done by human moderators (such as for ProMED) or automated. Most commonly, there are two types of classification approaches [131].

- **Keywords based classification**: documents are categorized based on the presence of predefined keywords.

- **Machine learning classification**: classifiers are trained on manually labeled data and automatically learn rules to label unclassified news articles (supervised methods), or use unsupervised methods that autonomously identify patterns and structures within data to classify documents that share similar characteristics.

Most EBS systems (PADI-web, HealthMap, MedISys) use binary classification categorizing news articles as either 'Relevant' or 'Irrelevant,' as shown in Figure 1.3, Step 2.
The first version of PADI-web used a keyword-based classification approach, where articles are classified as relevant if they contain in the text (title and body) one of the keywords related to an outbreak event (e.g. 'outbreak' 'cases' 'spread') [6].
MedISys classification relies on an approach involving Boolean combinations and keyword weightings. A document is considered relevant if it matches one of a predefined set of alerts. Two types of signals (i.e. single and combination) are implemented. A single signal consists of attributing positive and negative weights to relevant and irrelevant keywords. An article is kept if the sum of the keyword weights it contains is above a given threshold. A combination signal is based on keywords combined by Boolean expressions (i.e. 'AND' and 'AND NOT'). Documents are selected if they contain at least two relevant keywords and do not include any irrelevant keywords.
Systems like HealthMap, BioCaster, Argus, and GPHIN rely on supervised machine learning classifiers, specifically Bayesian algorithms and Support Vector Machines (SVMs).
HealthMap uses a Bayesian machine learning algorithm, relevant documents are then classified by location [57]. GPHIN computes a relevance score for each report, reflecting the SVM classifier's confidence. Expert moderators further verify classifications from GPHIN, HealthMap, and Argus. Articles with high relevance scores are kept, while low-scoring reports are automatically discarded. Experts review medium-relevance reports and check automatically discarded articles to ensure no relevant information is missed. BioCaster's classification is totally automated, relying on a naive Bayes classifier trained on a gold-standard corpus [35]. The second version of PADI-web 2.0 [161], integrated a machine learning classifier in addition to keyword-based methods. Later, [156] introduced PADI-web 3.0, which featured fine-grained classification of sentences to refine the notion of relevance and identify specific categories.
Modern EBS systems aim at improving the classification task beyond traditional machine learning approaches by focusing on Natural Language Processing (NLP) techniques [110]. For example, GRITS uses ensemble learning with logistic regression classifiers, where each classifier estimates the probability that a document is associated with a specific disease [79]. Additionally, the multilingual news surveillance system DANIEL leverages repetition and prominence (the beginning and the end of a news text often comprises the salient zones), in news writing, avoiding language-specific NLP toolkits by focusing on the general structure of journalistic texts [93]. Recent versions of these tools integrate language model approaches, such as the PADI-web for plant health surveillance [132].

Information extraction

Information extraction (IE) is the process of converting unstructured text into structured data containing information of interest. Different methods are used to achieve this task, including rule-based approaches and machine learning methods, as well as advanced models like transformers [118] and Bidirectional Encoder Representations from Transformers (BERT) [140, 98]. In the context of EBS, the objective is to extract relevant epidemiological information disseminated throughout the text, focusing on epidemiological data such as spatial and temporal entities (indicating where and when the event occurred), as well as thematic entities related to the host and causal agent (such as the pathogen or disease causing the event), as illustrated in Figure 1.4 [80].

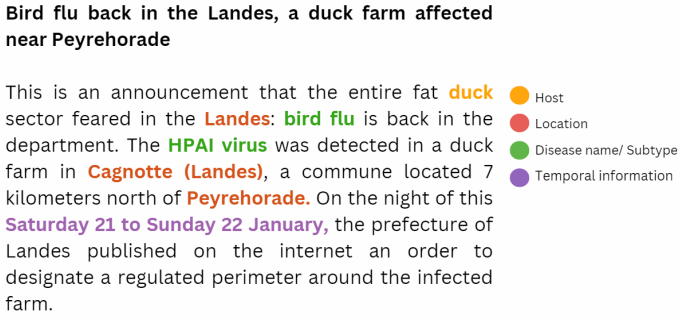


Figure 1.4: Snippet of an article detected by PADI-web, with highlighted epidemiological entities.

Several challenges are encountered in this step. The vocabulary used is diverse, especially when it relates to animal diseases, due to the existence of numerous hosts and a less formal vocabulary compared to humans for describing clinical signs [5]. Outbreaks' spatial information can be provided at different levels of granularity, and multiple events and locations may be mentioned within the same documents, adding to the ambiguity [159]. For example, a document might specify one location as the origin of the disease while describing the epidemiological situation in another location, or it could notify about an event in one location while detailing control measures in a neighboring area. Several studies have addressed the challenges associated with extracting the spatial information [149, 4, 155].
Not all EBS systems perform automated event extraction; for example, ProMED operates through a human-moderated process. In this case, experts extract and summarize key epidemiological information from reports before publishing them. For HealthMap, events extraction operates through an unsupervised approach. Where extraction of the epidemiological data relies on the document structure based on the hypothesis that the most relevant information appears at the beginning of a news report. Diseases and locations are first searched in the title, then in the document headlines, and finally in the full content. Experts further correct any errors in extractions when necessary [25].
In PADI-web, two types of approaches are employed and combined for entity extraction from

texts: (1) dictionary-based approaches and (2) classifier-based approaches. The dictionary-based approach entails matching terms from a document with a predefined list of keywords. Some dictionaries may include an ontological structure rather than a simple list of terms [159].
In the GRITS system, the information extraction process involves transforming words into vectors using the term frequency-inverse document frequency (TF-IDF) method. This pipeline begins with feature extraction through pattern-matching tools, which identify and extract relevant terms and phrases related to disease outbreaks, such as disease names, symptoms, locations, and dates. Once these features are extracted, the words are transformed into numerical vectors using the TF-IDF method, highlighting significant terms by evaluating their importance within the document relative to a larger corpus [93].

Communication

This step involves the communication of information to relevant authorities (such as national public health networks) or broader networks (including end-users of EBS systems). Various output are possible depending on the EBS system.
ProMED for example, publishes reports to the website ProMEDmail.org, as shown in Figure 1.5. In addition to sending e-mails to the subscribers [29]. HealthMap displays relevant

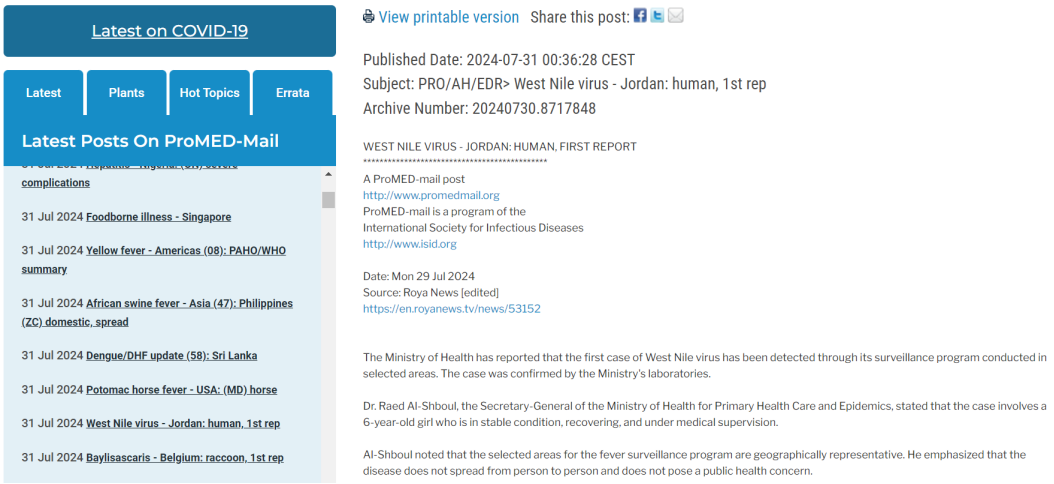


Figure 1.5: ProMED interface: Example of the latest published reports.

events on an interactive map accessible at Healthmap.org (see Figure 1.6). Its filtering and visualization features enable users to efficiently identify pertinent elements within their areas of interest, bringing structure to an otherwise overwhelming amount of information [57].
PADI-web offers several types of output. Users can navigate the platform to filter documents based on classification, date, and location (see Figure 1.7). The system also provides histograms to visualize the number of news articles over time, with aggregation options by day, month, or year, as well as a visualization feature per map. Moreover, users can export

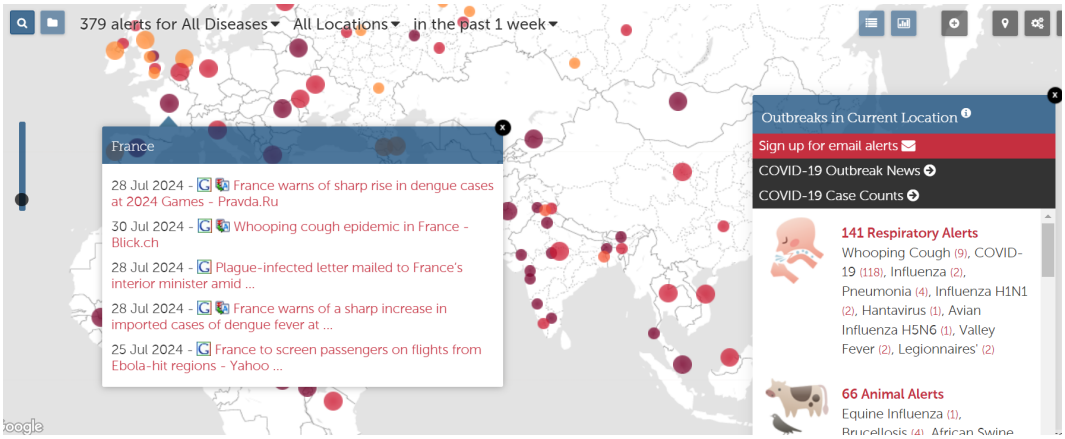


Figure 1.6: HealthMap interface: Display of recent outbreaks on the map and access to related articles.

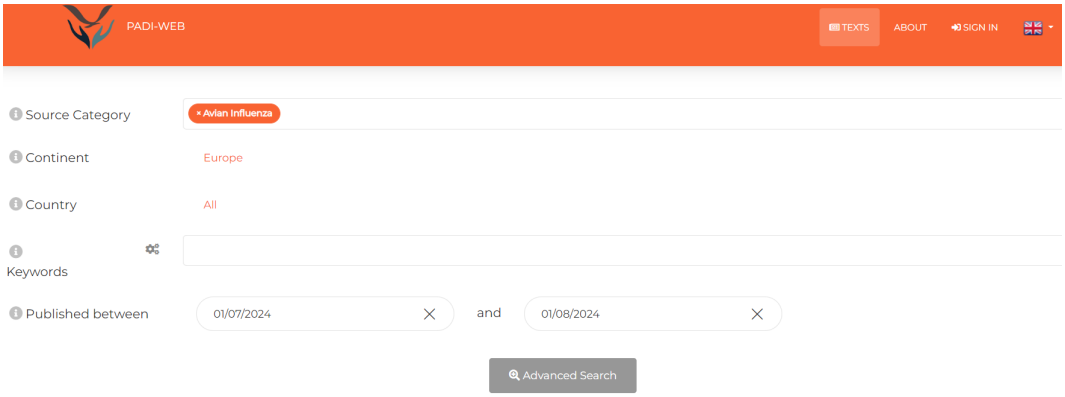


Figure 1.7: PADI-web interface: document filtering.

structured datasets containing extracted epidemiological entities from search results in formats such as CSV, JSON, and XLS. Additionally, the system includes an automatic email notification feature, that enables end-users to receive timely updates and alerts [160].

Effectiveness of EBS systems in disease surveillance

Event-Based Surveillance (EBS) systems have demonstrated their effectiveness in both human and animal health contexts.

For example, in the context of human health, ProMED played a crucial role during the 2003 SARS (Severe Acute Respiratory Syndrome) outbreak, sending daily reports that alerted medical staff at a hospital in Toronto. This early warning system allowed the hospital to isolate patients as soon as the disease appeared there [171]. GPHIN detected the SARS outbreak in southern China in 2003, issuing an early alert based on information from Chinese electronic media [115]. It also significantly contributed to monitoring the Ebola outbreak in

West Africa in 2014 [41]. During the 2009 H1N1 pandemic, HealthMap demonstrated impressive results with an average delay of just 12 days between the notification of suspected cases and their confirmation, which greatly enhanced the effectiveness of health responses [88]. In the 2020 COVID-19 pandemic, ProMED provided valuable information that helped identify new outbreak clusters [20].
In the field of animal health, in January 2014, ProMED was the first to report cases of African swine fever in wild boars in Lithuania, following alerts from local hunters [29]. A study on the 2019 avian influenza outbreak highlighted that PADI-web was very efficient in early detection of cases in wild birds [158].

Despite their effectiveness, EBS systems have some limitations, which will be discussed in the next section.

1.3.2 Limitations and discussion

While EBS systems have demonstrated their effectiveness in terms of relevance and early detection, several limitations can be highlighted.

First, the methods discussed in Section 1.3.1 involve machine learning and NLP techniques for document classification. Supervised methods are widely used and have shown satisfactory results in various studies for text classification [82]. However, these methods face several challenges: First, the success of supervised models heavily depends on the quality of annotated data. Annotating data can be expensive and time-consuming [114, 43]. In addition, balanced class distribution facilitates easier training and prevents bias. However, epidemiology-related datasets, like many real-world datasets, often exhibit imbalanced class distributions [114]. Moreover, supervised methods are limited to predefined categories used during training, and classification of epidemiology-related texts can be ambiguous, as some documents may mention a disease without reporting an outbreak, or may provide general disease information rather than specific outbreak details [157].
Keyword-based methods, such as dictionaries and ontologies, also face significant challenges. First, these methods require frequent updates to include new terms, which involves time-consuming manual work [159]. Second, adapting keyword-based methods to unknown diseases and evolving vocabulary is challenging, as they are based on predefined case studies. For example, a retrospective study conducted by [160] showed that during the early stages of the COVID-19 outbreak, keywords such as ‘pneumonia symptoms’ and ‘mystery illness’ were crucial for detecting relevant reports. However, once the disease was identified, the vocabulary shifted to terms related to the virus family and specific COVID-19 acronyms, highlighting the need for dynamic and adaptable keyword methods in EBS systems. Another major limitation is the lack of a formal definition of "relevance," which complicates the comparison of EBS system performance [12, 156]. For instance, in the PADI-web EBS system, the "Relevant" class includes articles related to outbreak declarations, preventive and control

measures, as well as economic and political consequences. In contrast, general information like economic and political consequences is considered irrelevant in MedISys and ProMED [96, 171].

Moreover, disease-related characteristics and environmental drivers can significantly influence how information is processed and reported. For example, West Nile virus outbreaks are more likely during the summer when vectors are active [66], increasing the likelihood of relevant events being reported during that period. However, to our knowledge, no EBS system currently incorporates environmental data. All the classification methods used to categorize articles rely solely on the epidemiological data found in the texts.
Each disease has a distinct outbreak definition, and its outbreak pattern often varies due to external risk factors. The spatio-temporal aspects of an epidemiological event play a crucial role in determining its 'relevance' [86]. This variability can reduce the robustness of classification models, especially when encountering unseen outbreak patterns that differ from those in the training data [111].

Based on the elements and limitations discussed in this chapter, our objective is to propose an unsupervised method that takes into account epidemiological data found in the documents (detected articles) and environmental data found in external sources. Several researchers have applied unsupervised machine learning algorithms to bio-informatics and text mining areas [163, 91, 94]. These methods help overcome the limitations associated with supervised models, as described in Section 1.3.2. For instance, [163] demonstrated the potential of these methods in identifying novel patterns and relationships from electronic health records without relying on manually annotated labels. [91] compared both supervised and unsupervised methods for biomedical text classification and found that unsupervised topic clustering methods are robust and applicable in real-world settings. [94] proposed an unsupervised machine learning model that detects latent infectious disease information from individuals' social media messages, using textual and temporal information along with sentiment analysis.
For the integration of environmental data, an effective approach is to rely on the results produced by risk-based surveillance methods and spatial modeling of infectious diseases. These methods are fundamental research fields that enhance public health preparedness and strategies for managing outbreaks [97, 64], and they rely on the spatial distribution of disease risks factors (environmental data) [77]. They help to highlight surveillance areas and adapt prevention and control measures when necessary [17]. This aspect of epidemiological surveillance will be explored in Chapter 2.

1.4 Conclusion

In this chapter, we have established the contextual framework for our work by detailing key components of EI framework, including IBS and EBS. We described main EBS systems highlighting their characteristics and limitations.
From this presentation, we have identified several challenges, some inherent to the methods of classification and data extraction in EBS systems including both supervised and unsupervised methods, and others related to the onset of diseases. These challenges vary in complexity depending on the disease type. Each type, such as vector-borne diseases, transboundary diseases with resistant viruses, or zoonotic diseases affecting multiple hosts, presents unique characteristics and therefore requires different surveillance approaches. A key observation made is that a significant amount of data is available and from various sources and of different natures (textual epidemiological data, spatial environmental data, etc.). Our hypothesis is that combining these data within the context of EBS can enhance the EBS systems in terms of classification and early detection as it places the detected events within their environmental context. To address these challenges, our objective is to propose an unsupervised method that integrates both epidemiological data from detected articles and environmental data from external sources. Unlike traditional text-based methods that classify each article individually, our approach evaluates the collective impact of all detected events within their environmental context to enhance classification accuracy.

For the advantages presented earlier in unsupervised methods, the goal of this work is to develop an unsupervised model that combines epidemiological and environmental data in an EBS context. We aim for the model to be robust, generic, and easily adaptable to various case studies. Therefore, we selected three distinct case studies representing different epidemiological systems: a zoonotic disease (Avian Influenza, AI), a transboundary disease (African Swine Fever, ASF), and a vector-borne disease (West-Nile virus Disease, WND).
Before building the method, the first step is to better understand the selected case studies, the relevant epidemiological features and the key environmental drivers related to each one. Thus, in the next chapter (Chapter 2), we will focus on these diseases, their characteristics, and their epidemiological surveillance context.

CHAPTER 2

CASE STUDIES: DISEASES CHARACTERISTICS AND EPIDEMIOLOGICAL CONTEXT

This chapter provides insight into three selected case studies; Avian Influenza (AI), African Swine Fever (ASF), and West Nile Disease (WND), that will be used to evaluate the robustness and genericity of our proposed model. Each of these diseases represents a different epidemiological system with distinct characteristics (transmission mechanisms, affected hosts, epidemiology and surveillance), which will be detailed in this chapter. All three are notifiable and have global coverage, ensuring comprehensive data availability from both IBS and EBS systems.

2.1 Introduction

A good starting point for understanding any infectious disease is to get familiar with its epidemiological triad. This model is fundamental in understanding disease patterns. According to it, the disease process arises from a complex interaction among factors associated with: 1. the pathogen, 2. the host, and 3. the environment [39, 84], as presented in Figure 2.1. What appears as very schematic and simple at first glance reveals multiple layers of complexity when considering the unique characteristics of each disease. For example, some diseases may have multiple transmission routes, such as vector-borne transmission and/or direct contact transmission. In addition, the term 'environment' is broadly used to encompass diverse factors and determinants of infectious diseases [39], including socioeconomic and demographic influences. Nevertheless, the interest of this model lies in that it compels us to approach each case study from an integrated perspective, which is particularly relevant in this study as we aim to build a model that combines epidemiological and environmental data in the context of EBS.

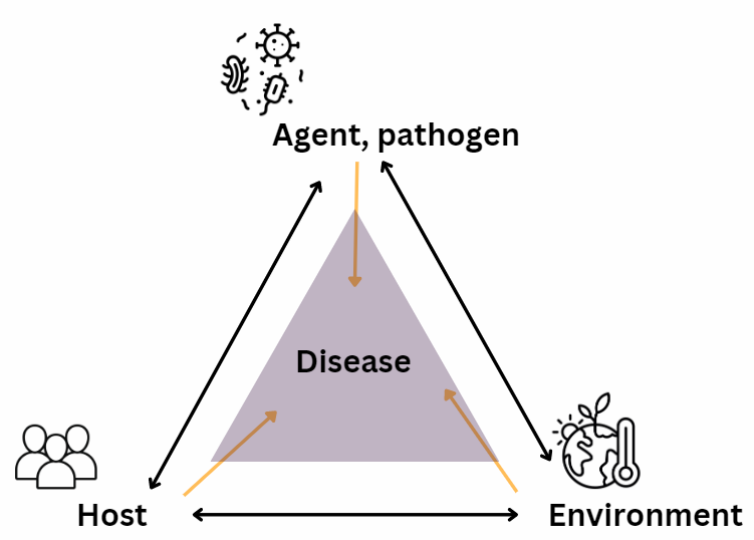


Figure 2.1: Epidemiological Triad.

The motivation behind selecting these diseases lies in their diverse characteristics, including variation in transmission patterns, pathogens, affected hosts, surveillance challenges and environmental drivers. These differences make them ideal candidates for evaluating the genericity and robustness of our model which is designed to be independent from a specific disease or a host.

Avian Influenza (AI), African Swine Fever (ASF), and West Nile Virus (WND) were selected because they are globally widespread and receive significant media coverage, which enhances data availability in EBS systems. Additionally, they address particularly topical

issues in Europe, with significant public health and economic impacts. They are also subject to notifications to the WOAH and the WHO, ensuring the availability of relevant and confirmed event data in IBS systems.
This chapter aims to present the main characteristics of AI, ASF and WND as well as their epidemiological context, with the goal of better understanding which epidemiological features and environmental drivers to consider.

2.2 Avian Influenza

2.2.1 Disease characteristics

Avian Influenza (AI), also known as bird flu, is a highly contagious viral disease that primarily affects avian species, both domestic and wild [28]. AI viruses possess a great zoonotic potential as they are able to infect different avian and mammalian animals hosts, from which they can be transmitted to humans [112]. AI viruses are typed according to their pathogenicity in poultry into highly pathogenic avian influenza (HPAI) with flock mortality as high as 100%. These viruses have been restricted to subtypes H5 and H7, although not all H5 and H7 viruses cause HPAI [28], and low pathogenic avian influenza (LPAI) that cause a milder, primarily respiratory, disease.
Aquaric birds belonging to the orders Anseriformes (ducks, geese) and Charadriiformes (shorebirds) act as natural reservoirs of AI viruses (see Figure 2.2). For this reason, the proximity to water is described as a significant risk factor for virus transmission [85]. Ducks, geese and wild water fowl, suffer mild illness whereas poultry birds are more severely affected and are responsible for the large outbreaks and epidemics in poultry (see Figure 2.2). AI spread worldwide, via migratory birds and poultry trade activities.
The symptoms of AI in birds can vary significantly depending on the virus's pathogenicity. LPAI infections may present subtle signs such as ruffled feathers, slight weight loss, transient reductions in egg production, and mild respiratory symptoms. In contrast, HPAI can cause severe disease, particularly in chickens and turkeys, with a sudden onset of severe symptoms such as diarrhea, edema, nervous symptoms, and a rapid cessation of egg production, often leading to mortality within 48 hours [60]. When humans are infected with AI viruses, they often experience no symptoms or mild symptoms such as cough, headache, weakness, and runny nose. However, some strains can cause severe disease, particularly among infants and individuals with underlying medical conditions, potentially leading to pulmonary inflammation and death [9].

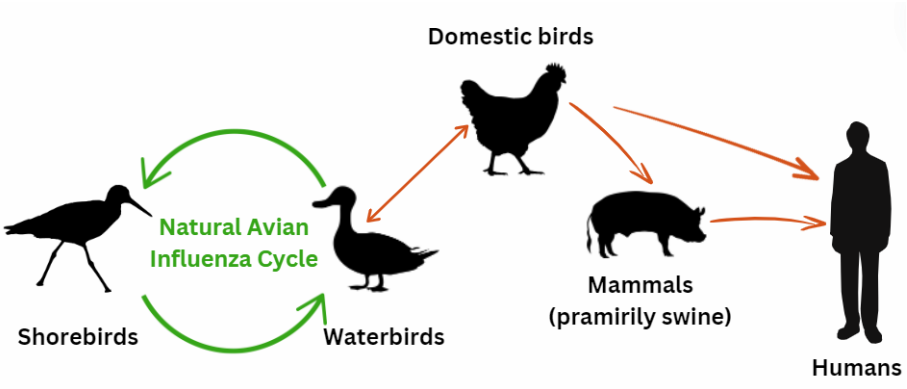


Figure 2.2: Possible routes of transmission of AI, natural cycle (green arrows), other transmissions (orange arrows).

2.2.2 Epidemiology and surveillance

Avian influenza outbreaks have displayed a diverse geographical spread and impact over the years. Initially, AI was predominantly observed in poultry population across Asia where, the HPAI virus is endemic in many countries [73]. In the beginning of the 2000s, many aspects of the epidemiology of AI infections in poultry and other birds appear to have changed dramatically from those established in the preceding century. The number of outbreaks of the HPAI disease has increased alarmingly in the last 10 years and (see Figure 2.3), even more noticeably, the impact in terms of the number of birds involved and the costs of disease control have dramatically escalated.

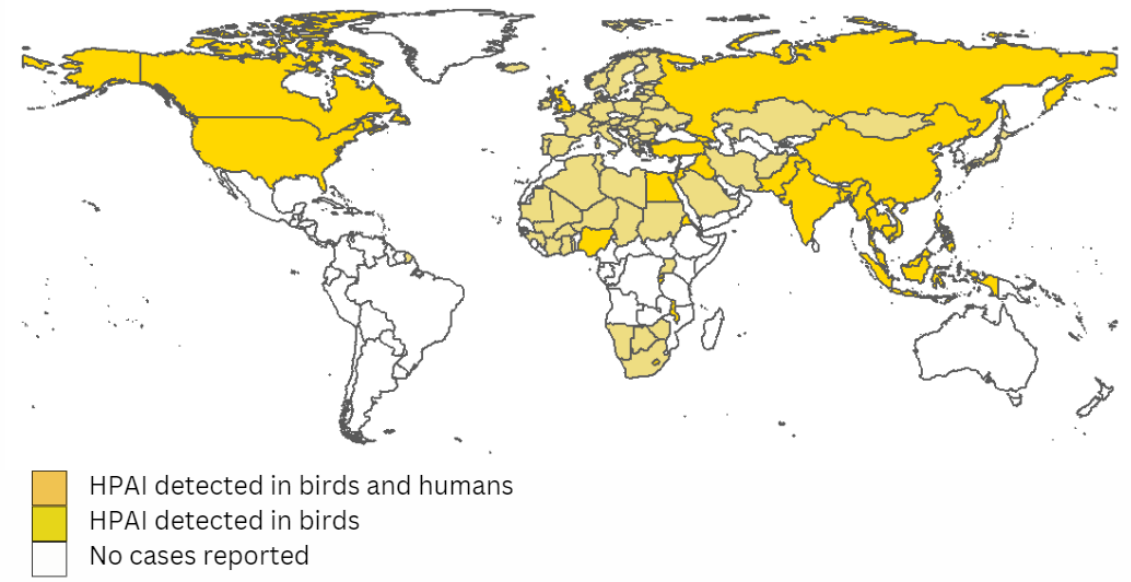


Figure 2.3: Global distribtuion of HPAI in 2023, adapted from [42].

Currently, global avian influenza surveillance primarily involves wild birds, poultry, re-

lated environments, human cases, and high-risk populations. The surveillance content and methods differ accordingly [44]. In wild birds, AI surveillance relies on three primary strategies. Active surveillance that involves capturing birds or hunting waterfowl in specific habitats or high-risk areas with known outbreaks. Passive surveillance that focuses on examining unusual bird deaths and visible signs of illness. Finally, sentinel surveillance that monitors domestic ducks in typical habitats to track disease spread among wild birds. Collected samples, such as cloacal swabs, feces, and environmental samples, are analyzed for serological and pathogenic indicators [150]

In domestic birds, the targets for surveillance primarily encompass domestic fowl, waterfowl, and ornamental birds, with a particular focus on ducks due to their proximity to migratory waterfowl. Surveillance methodologies include clinical sign observation and laboratory examination. On poultry farms, surveillance can be achieved by monitoring signs associated with AI infection.

When AI is detected in a given area, a surveillance zone is established to contain and monitor the outbreak. Control measures include closing all poultry and egg markets within a 10-km radius of the infected location, installing infected area sign-boards within a 3-km radius, and establishing a surveillance perimeter with a radius of 3–10 km in areas where no vaccination is implemented (Figure 2.4). The zone is considered free from AI when no cases are reported 3 to 4 weeks after the outbreak's detection [9, 85, 127].

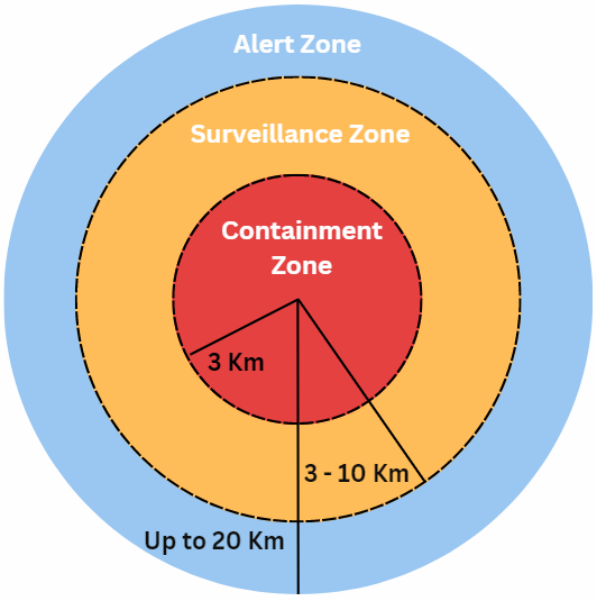


Figure 2.4: Schematic representation of delimitation zones in an infected area adapted from [127, 162]

The surveillance of human infection with AI viruses primarily depends on medical reporting in healthcare facilities. Moreover, active surveillance of high-risk occupational populations is conducted in several countries. For instance, China regularly performs active surveillance among high-risk groups, such as poultry farmers, poultry traders, and poultry

slaughtering and processing staff. Different countries have adopted various strategies to control HPAI. In Europe and North America, the approach typically involves culling infected and suspected birds. In contrast, some countries, such as China, have adopted a strategy that combines culling with vaccination

On both global and regional levels, the AI surveillance network has shown considerable growth [44]. Various networks have been established to enhance early detection and response to AI outbreaks and facilitate communication and data exchange. Examples include the Global Influenza Surveillance and Response System (GISRS) [75] and the OFFLU network [37]. Internationally, the World Organisation for Animal Health (WOAH) mandates member countries to promptly report outbreaks of HPAI in domestic and wild birds, as well as LPAI subtypes H5 and H7 in poultry, and any unusual mortality events among wild birds [150]. Additionally, human cases of avian influenza must be reported to the WHO. Confirmed cases of avian influenza are documented in various IBS databases mentioned in Chapter 1, such as EMPRES-i FAO and WAHIS-WOAH. These databases provide comprehensive information, including the location of outbreaks, observation and confirmation dates, and other relevant details such as the subtype, host, and number of cases.

When it comes to event-based surveillance (EBS) of AI, this disease benefits from extensive media coverage. Both general and specialized media outlets actively report on various aspects of AI, including outbreaks, control measures, and the introduction of new virus strains. As an example, in a study conducted by [158], websites, such as: Poultry Site[1] and WATTpoultry[2], appeared to be valuable sources for both PADI-web and HealthMap EBS systems.

2.3 African Swine Fever

2.3.1 Disease characteristics

African Swine Fever (ASF) is a very complex viral disease that affects only porcine species (both wild and domestic), producing a variety of clinical signs and lesions from acute, subacute and chronic. It can easily be confused with classical swine fever (hog cholera), or other hemorrhagic diseases, for this reason laboratory test is required to establish a correct diagnosis [3].

ASF transmission can occur directly through close contact with infectious animals or indirectly by ingesting infected pork products, touching contaminated objects (fomites), or possibly via mechanical vectors. Additionally, ASF can be effectively transmitted by the

[1] Poultry Site is a specialized website providing news and information on the poultry industry. Available at https://www.poultrysite.com/

[2] WATTpoultry is a specialized website providing industry news and analysis on poultry and avian diseases. available at https://www.wattagnet.com/

biological soft tick vector, genus *Ornithodoros* spp., if present. However, *Ornithodoros* spp. is not considered significant in the current ASF epidemic in Central and Eastern Europe. In the absence of this tick vector, the most efficient transmission method is direct contact with the blood of infected animals (see Figure 2.5).

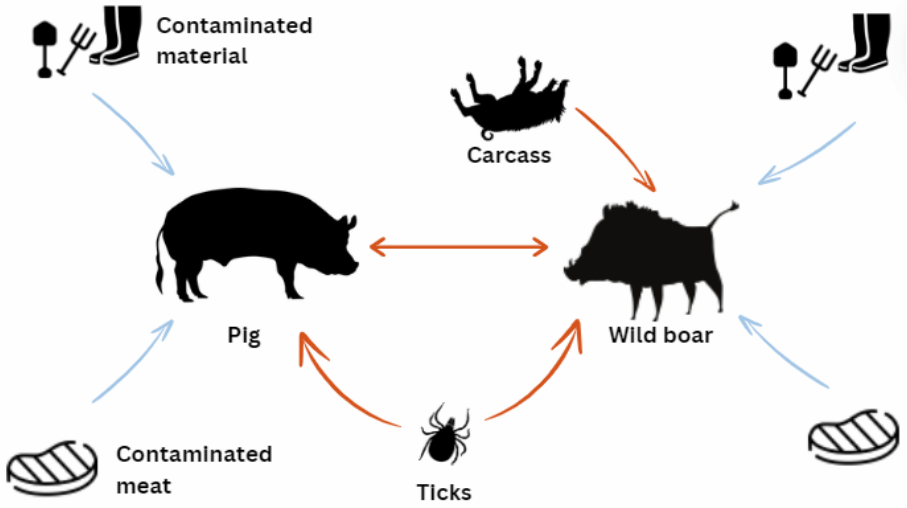


Figure 2.5: Possible routes of transmission of ASF, biological transmission (orange arrows), and mechanical transmission (blue arrows).

Several factors make the ASF virus a serious threat to the global swine industry and national economies. Mortality rates can reach 100% in acute cases. The high morbidity and mortality associated with the virus, alongside the absence of an effective vaccine and its high resistance, contribute to its severity. In addition, ASF virus is very stable in the environment and in food, able to remain in preserved meat for up to 6 months, which allows for possible transmission, especially to wild boars, through abandoned human food waste [81].

2.3.2 Epidemiology and surveillance

ASF is endemic in sub-Saharan African countries. Eradicated from Western Europe since the late 1990s, except for an endemic form in Sardinia, ASF was reintroduced to Georgia in June 2007. The introduced strain was identified as closely related to strains found in East Africa and Madagascar. The most probable hypothesis for the virus's introduction to Eurasia suggests contaminated pork products from a cruise ship. Through the trade of pigs and pork, the virus rapidly spread among domestic pig populations and wild boars populations.
Azerbaijan and Russia were affected in 2008, impacting both wild and domesticated populations. Despite Russian authorities' interventions in 2009, the virus was detected in pigs near the European border north of Saint Petersburg, close to the Estonia-Finland border. In 2012, Ukraine and Belarus reported their first outbreaks, followed by the virus's first introduction to the Baltic countries (Latvia, Lithuania, Estonia, and Poland) in 2014 [61]. During that

period, ASF infections were primarily observed on pig farms with low biosecurity, occasionally spilling over into the wild boar population. It was initially predicted that the disease would fade out in wild boars once controlled in domestic pigs due to the high mortality rate and lack of long-term carriers. However, this proved incorrect in Poland and the Baltic states, where ASF persisted in the wild boar population independently of outbreaks in domestic pigs [34, 119].

In 2023, African swine fever (ASF) impacted 14 European countries, notably Croatia and Sweden where ASF emerged among wild boars, and Greece where ASF re-emerged after being free since 2021. The number of ASF outbreaks among domestic pigs in the EU was five times higher than in 2022, reaching a similar magnitude to that in 2019 [47]. The global distribution of ASF is shown in Figure 2.6. ASF surveillance in wild boar is carried out ei-

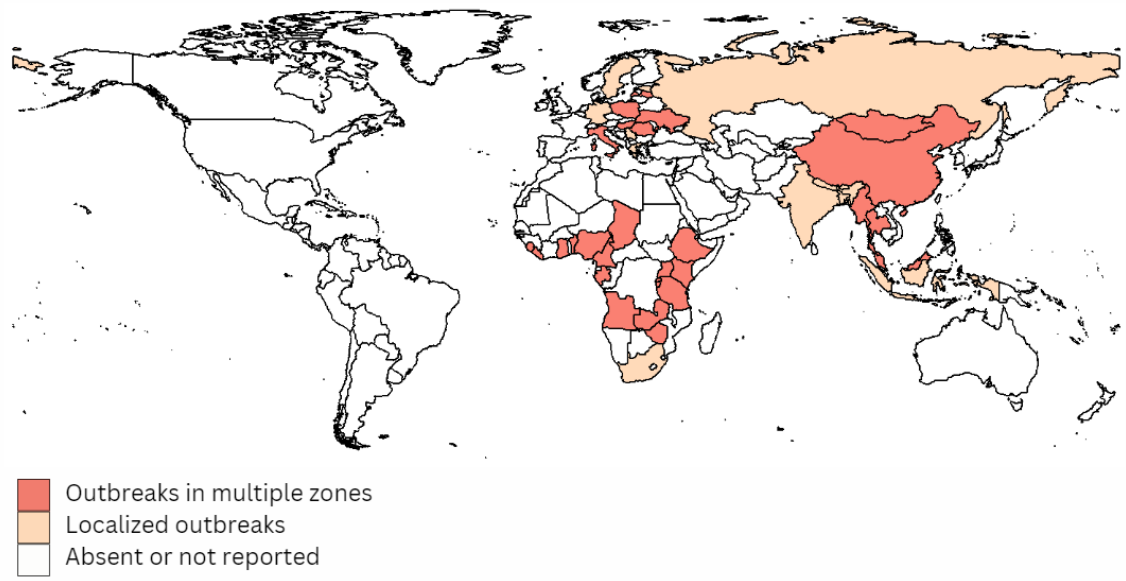


Figure 2.6: Global distribution of ASF in 2023, adapted from [165]

ther by testing all the wild boar found sick or dead for virus detection (passive surveillance) or by testing for virus (and antibodies) all hunted wild boar (active surveillance) [63]. ASF surveillance focuses on both passive and active monitoring of wild boars and domestic pigs. The main strategic aims of surveillance in domestic pigs are early detection of potentially infected holdings [162], which is challenging because of the wide range of non-specific clinical signs produced[56]. In livestock, if the disease is suspected (when pigs display clinical signs, show post-mortem lesions...etc), the holding must be placed under official surveillance until the ASF situation is clarified through laboratory tests. Key measures include: counting all pigs by category, compiling lists of sick, dead, or potentially infected pigs, and creating a map of the holding for epidemiological investigations. All pigs should be confined to their living quarters, and no pigs or pig products should leave the holding until ASF is ruled out. Additionally, movement of people and vehicles to and from the farm should be restricted, and disinfection protocols should be enforced at stable entrances and exits [162, 63].

Upon ASF confirmation, immediate actions must be taken in the affected holding. All pigs must be euthanized without delay, and samples collected for further epidemiological investigation, particularly to trace the virus's introduction and estimate how long ASF may have been present before notification. The investigation should also determine the virus's possible origin, identify contact holdings potentially infected from the same source, and assess whether vectors (e.g. soft ticks) or wild boars contributed to the infection. Any pigs, meat products, semen, ova, or embryos that left the holding should be traced, and thorough cleaning and disinfection of the holding should be performed [162].
Being a highly contagious disease, like HPAI, the establishment of a protection (containment) zone with at least a 3 km radius and a surveillance zone with at least a 10 km radius around the outbreak site is required (see Figure 2.4). Movement and transport of pigs are prohibited, and restrictions cannot be lifted earlier than 30 days post-cleaning in the protection zone and 20 days in the surveillance zone. Due to the resistant nature of ASF virus, restocking is permitted no sooner than 40 days after cleaning and disinfection.
ASF is a notifiable disease and its notification to the WOHA is mandatory [58].
In the context of EBS, numerous media outlets, both specialized (such as PigSite[3]) and non-specialized, publish reports on ASF outbreaks, virus introductions, and control measures. Which are collected daily by EBS systems, such as PADI-web.

2.4 West-Nile virus disease

2.4.1 Disease characteristics

West-Nile Disease (WND) is a multi-host mosquito borne virus belonging to the Japanese encephalitis (JE) antigenic complex (genus *Flavivirus*, family *Flaviviridae)* [102]. The most common route of WND infection to humans is through the bite of an infected mosquito of the genus *Culex*. Mosquitoes become infected when they feed on infected birds that have high levels of WND virus in their blood. This cycle of transmission between birds and mosquitoes is referred to as "enzootic amplification". Infected mosquitoes can then transmit the virus when they feed on humans or other animals. People, horses, and most other mammals do not develop high-level viremia[4], they do not contribute to the transmission cycle and hence are traditionally called "dead-end" hosts [136]. The transmission of the virus from mosquitoes to humans or horses is known as spillover, referring to the infection of unintended hosts outside the primary enzootic cycle (see Figure 2.7).

While most human infections with WND, around 80%, are asymptomatic and often go unnoticed, approximately 20% of individuals develop flu-like symptoms known as West

[3]PigSite is a specialized online platform for swine diseases and industry news. Available at https://www.thepigsite.com/
[4]Viremia refers to the presence of viruses in the blood.

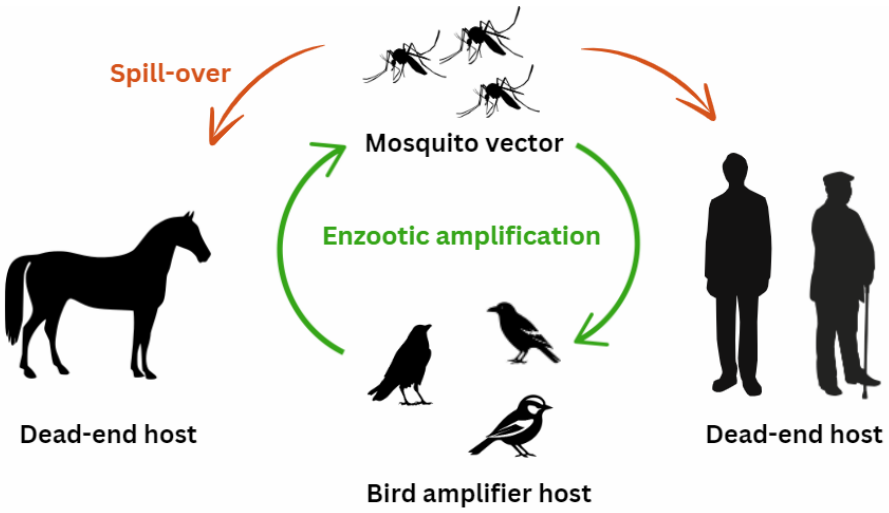


Figure 2.7: Transmission pathways of WND virus, enzootic cycle (green arrows), other transmissions (orange arrows).

Nile fever. Additionally, about 1% of cases develop a severe, potentially fatal, neuroinvasive disease.

2.4.2 Epidemiology and surveillance

Since its first discovery in 1937 in the West Nile district of Uganda, the WND virus has undergone a substantial geographical migration and spread around the globe [52]. Historically, WND has been associated with asymptomatic infections and sporadic disease outbreaks in humans and horses in Africa, India, and Middle East [19, 125]. However, starting in the mid-1990s, the frequency, severity, and geographic range of WND outbreaks increased dramatically, and the virus has caused frequent outbreaks of severe neuroinvasive disease in humans and horses in Europe and the Mediterranean Basin [19, 31].
In 1999, The virus reached the American continent, marked by a cluster of encephalitis cases reported in the metropolitan area of New York. Within three years, the virus had spread the contiguous U.S. and the neighboring countries of Canada and Mexico [125]. Since its discovery, the virus has spread to a vast region of the globe and is now considered the most important causative agent of viral encephalitis worldwide [52, 19].

WND surveillance like other diseases discussed relies on a multi-disciplinary approach, involving experts from animal, human, and environmental health. However, being a mosquito-borne disease, it is highly influenced by environmental data and seasonal patterns. The main objective of WND surveillance is the early detection of virus circulation among birds (particularly corvids), mosquitoes, horses, and humans at a local level. Thus, each country maintains its own national and regional surveillance systems, which are adjusted to reflect current epidemiological and ecological conditions [103].

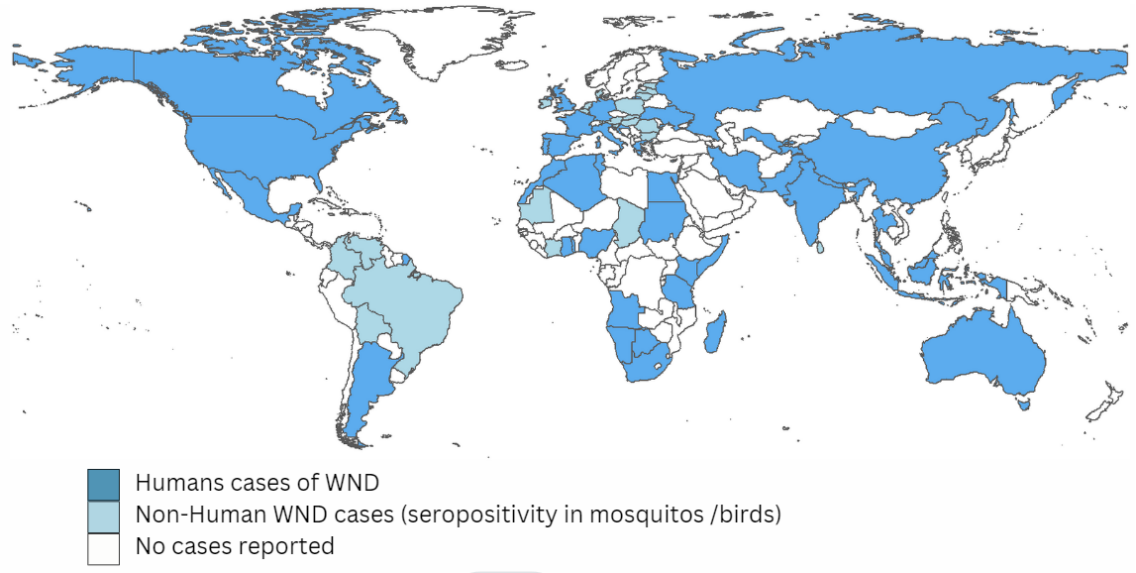


Figure 2.8: Global distribution of WND in 2024, adapted from [31, 51].

For example, in the United-States, where WNV is endemic, ArboNET is a surveillance system that was implemented by the CDC in the 2000s to monitor cases in humans, birds, and horses. ArboNET maintains data on arboviral infections among presumptive viremic blood donors, veterinary disease cases, mosquitoes, dead birds, and sentinel animals. Since its creation that was initially targeting WND, ArboNET now also monitors other arboviroses such as dengue and zika [108].

At the European level, WND infection is notifiable for humans and equids. Although humans are considered a dead-end host as discussed in Section 2.4.1, one of the main goals of human WND surveillance is to prevent human-to-human transmission via the donation of contaminated substances of human origin, such as blood. Human cases are reported according to the EU case definition by national public health authorities and are recorded in The European Surveillance System (TESSy) at the European Centre for Disease Prevention and Control (ECDC) [169].

In addition to these surveillance systems, several studies have explored environmental variables and risk factors, including vegetation indices, temperature fluctuations, and socioeconomic factors, as disease drivers to better predict WND outbreaks and target surveillance areas [84, 14]. Currently, several equine vaccines for WND virus have been licensed and are successful in horses. However, there is no WND virus vaccine approved for human use, although several candidates have shown promising results in clinical trials [137].

2.5 Spatial modelling and risk-based surveillance

Besides the surveillance strategies discussed earlier in this chapter, the spatial modeling of infectious diseases [64] and risk mapping methods are two fundamental research fields that have been used to enhance public health preparedness and the planning of strategies to contain outbreaks [97].
The purpose of spatial modeling in animal and public health is threefold: to describe existing spatial patterns (descriptive) [76, 18], to understand the biological mechanisms that lead to disease occurrence (explanatory), and to predict future occurrences in different geographical areas or over time (predictive).
These methods rely on the spatial distribution of disease risks factors [77], and they help to highlight surveillance areas and adapt prevention and control measures when necessary [17].

Covariate group	**Risk factor**	**AI**	**ASF**	**WND**
Climate	Temperature	X		X
	Humidity	X		X
	Precipitation	X		X
	Frost			
	Snow			
Water	Surface water (wetlands)	X		X
Hosts	Domestic birds abundance	X		X
	Wild bird abundance	X		X
	Bird migration			
	Pig density			X
	Wild boar density			X
	Human population	X		X
Vector	Ticks		X	
	Mosquitos (*Culex* spp., *Aedes* spp.)			X
Agriculture	Farm densities			
	Farm location	X	X	
	Biosecurity measures	X	X	
	Wildlife Trade and Wildlife market	X		

Table 2.1: Risk factors associated with the diseases: AI, ASF, and WND.

The risk factors driving a given disease can be numerous, including environmental, climatic, socioeconomic, and demographic factors. Their importance varies depending on the characteristics of the disease. For instance, in mosquito-borne diseases such as malaria, dengue, and WND, environmental factors that create suitable conditions for the vectors such as warm temperature and humidity play a key role [90]. For zoonotic disease such as AI, several studies show that environmental drivers such as the proximity to wetlands, and poultry density are crucial [53]. For transboundary and highly contagious diseases such as ASF, factors such as trade movements, biosecurity measures, and reservoir populations are the

primary drivers [167].
Various methods can be found in the literature that aim at identifying the weights of disease risk factors. Multi-Criteria Decision Analysis (MCDA) is commonly used for this purpose, allowing experts to systematically evaluate and prioritize criteria based on their relative importance [143, 144, 124]. The outputs of these methods, such as risk maps and suitability maps, can serve as valuable sources for environmental data. Table 2.1 summarizes the disease risk factors that are associated with the three diseases:

2.6 Conclusion

In this chapter, we described three different diseases, detailing their characteristics, including symptoms, hosts, transmission cycles, as well as the surveillance and control measures employed to monitor them.
The literature review highlights that: first, surveillance strategies must be adapted to the unique characteristics and environmental drivers of each disease, which underscores the need of integrating expert knowledge in the surveillance methods. Second, it points out ongoing challenges in monitoring these diseases.
The three case studies presented here involve notifiable diseases that are monitored by various networks and IBS systems. Despite these efforts, challenges persist due to incomplete reporting of outbreaks, which may result from inadequate surveillance systems, or concerns about political and economic impacts [150] (especially for AI and ASF). Additionally, symptoms of these diseases can be non-specific and overlap with other illnesses, which can delay the monitoring process and notification.
These issues underscore the importance of EBS systems, which address the gaps in traditional surveillance and enhance the early detection and response to such diseases. Considering the elements presented in this chapter, and building on the limitations inherent to EBS systems discussed in Chapter 1, Section 1.3.2, the next chapter (Chapter 3), will present the intuition behind our proposed model that aims to integrate data-driven approach and expert knowledge to enhance EBS systems.

CHAPTER 3

FROM BIOLOGICAL INSIGHT TO COMPUTATIONAL DESIGN

In this chapter, we explain the intuition and inspiration behind our model. We begin by introducing the general concept of Artificial Immune Systems (AIS), a family of algorithms inspired by the human immune system. Next, we explore the danger theory, an immunological concept that served as the foundation for the Dendritic Cells Algorithm (DCA), which, as we will discuss in the following chapter, forms the basis for our model.

3.1 Introduction

Computer science have a great history of drawing inspirations from nature's designs; genetics and natural evolution has inspired the genetic algorithms [130]. The human brain has inspired the neural network model, which is one of the foundations of artificial intelligence [168]. Similarly, the human immune system has inspired the development of a family of algorithms called Artificial Immune Systems (AIS).

The immune system can be visualized as a series of defensive layers protecting the host. Once an antigen[1] enters the body, it faces two subsystems: the innate and acquired immune systems. These subsystems are interconnected and comprised of many types of cells and molecules produced by specialized organs [129].

The innate immune system is the body's first line of defense against intruders. It responds in the same way to all germs and foreign organisms, which is why it is referred to as the "non-specific" immune system [83]. While the adaptive immune system responds to previously unknown antigens, and builds a response to them that can remain in the body over a long period of time.

Classically, immunology has focused on the body's capacity to discriminate between antigens belonging to 'self' or 'non-self'. This foundational theory, introduced by Paul Ehrlich in 1891, has guided immunological research since its conception [62]. However, Numerous questions remained unanswered with this paradigm. For example, why do intestines contain millions of bacteria, yet the immune system does not react against these colonies of non-self invaders? [67]. Many antigens that enter the body are harmless, and it would be unnecessary and potentially harmful to trigger adaptive immune responses against them. Allergic conditions are good examples of deleterious adaptive immune responses against apparently harmless molecules [30]. In 1994, immunologist Polly Matzinger controversially postulated that the immune system's objective is not to discriminate between self and non-self, but to react to signs of damage to the body. This theory is known as the Danger Theory [105], and suggests that the immune system responds to the presence of molecules known as danger signals, In other terms it is the environmental context in which the antigens are perceived that will condition the immune response.

When considered from a computational point-of-view, the immune system can be seen as a rich source of inspiration, as it displays learning, adaptability, and robustness [153]. This remarkable information-processing biological system has caught the attention of computer scientists leading to the development of a family of algorithms called: Artificial Immune Systems (AIS) [1].

[1] An antigen is a large protein molecule capable of inducing an immune response in the body by the production of antibodies.

3.2 Artificial Immune Systems

AIS in the literature can be broadly categorized into two generations. The first generation relies on simplified immune models [87]. These algorithms draw inspiration from basic immunological concepts. For instance, the clonal selection algorithm, inspired by Burnet's clonal selection theory [27], suggests that only cells capable of recognizing an antigen will proliferate. Initially proposed to solve pattern recognition problems, it involves processes such as initial population generation, selection, cloning, hypermutation, and receptor editing [74]. Another example is the negative selection algorithm, which mimics the immune system's ability to distinguish between self and non-self. It is mainly used for anomaly detection and is beneficial for one-class classifications, outlier detection, and fault and intrusion detection problems [72]. However, these algorithms have often shown considerable limitations when applied to complex realistic applications. To address these challenges, a second generation has emerged that is more sophisticated and relies on interdisciplinary collaboration to develop a deeper understanding of the immune system, thereby producing more complex models. These new models draw inspiration from cutting-edge immunology, such as the Dendritic Cell Algorithm (DCA) [67], which is based on the Danger Theory [67]. In the next section, we will explain in detail the Danger Theory and the functioning of the DCA, as they form the basis of our model.

3.3 Danger Theory

3.3.1 Core concepts

The danger theory states that the recognition of an antigen by a cell is not due to the cell distinguishing between self and non-self, but rather depends on the environmental context (signals) in which the antigen is identified. This theory is based on the functioning of dendritic immune cells (DCs) [105], which form part of the body's first line of defence against invaders.

3.3.2 Dendritic Cells

DCs are a type of antigen-presenting cells. They are seen as detectors responsible for policing different tissues. They have the ability to combine a multitude of molecular information and to interpret this conflicting information for the immune system, which leads to the induction of responses against perceived pathogenic threats [32].
Throughout their lifespan, (DCs) exist in one of three states, namely "immature", "semi-mature", and "mature". In their initial state, the (DCs) are "immature". Then, based on the

concentration and the type of signals they are exposed to, (DCs) differentiate into either a "semi-mature" form to suppress the immune alarm, or a "mature" form to activate it [70] (see Figure 3.1).
In immunology, there are four main types of signals, namely the pathogenic associated molecular patterns (PAMPs), danger signals (Ds), safe signals (Ss), and inflammatory cytokines (I) [32].

- **Pathogen-associated molecular patterns (PAMPs)** are molecules found in groups of pathogens. These small molecular motifs are conserved within classes of microbes and can be detected by dendritic cells (DCs), leading to immune activation. The presence of PAMPs clearly indicates an abnormal situation.
- **Danger signals (Ds)** are released due to unprogrammed cell death, such as necrosis, which is caused by external factors like infection, toxins, or trauma. Ds indicate abnormality but with a lower confidence level compared to PAMP signals.
- **Safe signals (Ss)** are produced via the process of normal cell death, i.e. apoptosis. Ss are indicators of normality which means that the antigen collected by the DC, within this context, it is not harmful and the situation does not require an immune reaction.
- **Inflammatory cytokines (I)** are signals proving that there is an increase in temperature in the affected tissue. Inflammation signals have the effect of amplifying the other three categories of signals, but they have no efficiency when they are present alone in the system.

The robustness of the immune system mostly lies in the ability of the DCs to sense an early death cell (viewed as an outbreak event). The danger theory offers a multivariate detection approach that does not require a training phase [111]. This theory, combined with the behaviour of DCs, inspired the development of the DCA [32], a classification algorithm that has been successfully applied to a wide range of challenging real-world applications.

3.4 Dendritic Cells Algorithm (DCA) and related work

3.4.1 Original DCA version

DCA is a population-based system with each agent in the system is represented as a cell (DC). Each cell has the ability to collect data items called antigens, which represent the specific data instances that need to be classified [32]. The DCA was initially designed by [67] to be used as an anomaly detection algorithm. It consists of a four-phase process: Pre-processing and Categorization, Detection, Context assessment, and Classification (see Figure 3.2).

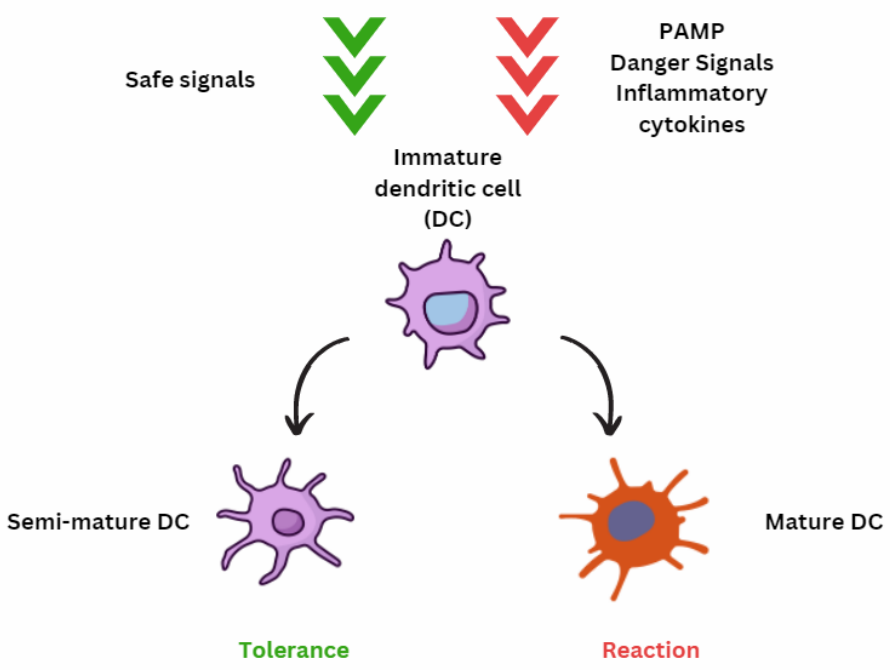


Figure 3.1: Maturation of the dendritic cells.

3.4.1.1 Pre-Processing and Categorization phase

In the pre-processing and categorization step, the most important features from the database are selected. Each feature is then transformed into a numerical value and assigned to an appropriate signal category. In other words, this process quantifies each feature and classifies it into the relevant category, allowing us to interpret the data as signals (Figure 3.2, phase 1).

- PAMPs: increase in proportion to the presence of data representing an "abnormal" situation, it is a confidence indicator of anomaly.
- Danger signals (*danger signal*): increase in proportion to the presence of data representing an "abnormal" situation, has lower confidence than PAMP signal (i.e. a lower weight).
- Safe signals (*safe signal*): increase in proportion to the presence of data representing a "normal" situation and has a negative weight.
- Inflamation signals: Amplify the output values of the other three signal categories. However, when present alone, they do not impact the state of a dendritic cell.

To perform data pre-processing, some DCA studies involve users or experts to select or extract the most interesting features and assign them to their appropriate [32] signal categories. Other DCA studies apply some dimensional reduction techniques such as the principal component analysis [70].

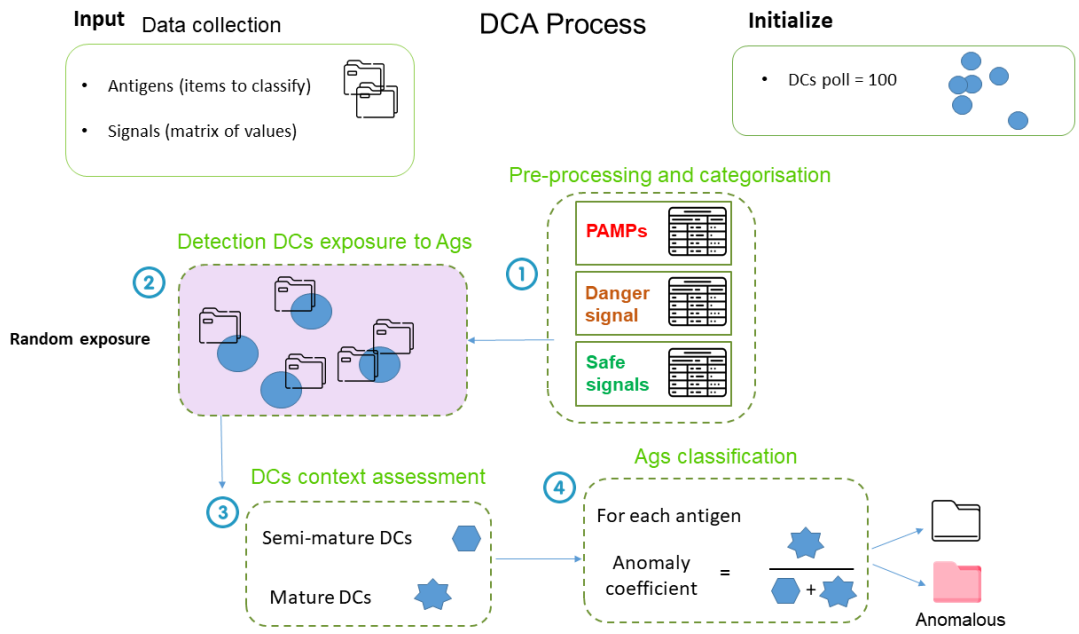


Figure 3.2: Representation of the DCA phases.

3.4.1.2 Detection phase

During the detection phase, each DC is exposed randomly to i Antigens (Ags), (Figure 3.2, phase 2). Based on the induced signal database, the algorithm processes its input signals to get three cumulative output signal values known as:

- The co-stimulatory molecule signal value (CSM)
- The semi-mature signal value (smDC)
- The mature signal value (mDC)

These cumulative output signal values are calculated using the defined input signal values and a set of weights, as shown in equation 3.1 where $C \in \{CSM, smDC, mDC\}$:

$$C[CSM, smDC, mDC] = ((W_{PAMP} \times PAMP) + (W_{Ds} \times Ds) + (W_{Ss} \times Ss)) \times 1 + I \quad (3.1)$$

where $PAMP$, S_S and S_D represent PAMP, safe and danger signals' values, and W_S and W_D their corresponding weights. PAMP Signals (PAMPs), Danger signals (Ds) have a positive weight and Safe signals (Ss) have a negative weight. The weight values can be either derived empirically from the data or from user-defined values [32]. I represents the inflammation signal. Although inflammation signals have occasionally been incorporated in various DCA implementations found in the literature, most proposed DCAs tend to ignore

these signals, focusing instead on PAMPs, DSs, and SSs [32]. Among the few studies that have incorporated this signal, we can cite [69].

These three DC output signals has two roles: first, it allows an assessment of the cell's context and the classification of the cell as semi-mature or mature; second, it is used to stop antigens sampling [55]. To limit the time spent sampling data, a migration threshold value (MT) is assigned to each DC in the population upon its creation. If a DC's CSM value exceeds the MT, the DC's exposure to Ags is stopped, otherwise the algorithm continues sampling and keeps calculating and updating the CSM values [33].

3.4.1.3 Context assessment phase

In the context assessment phase, upon migration, the cumulative output signals are evaluated, and the cell context is determined by selecting the greater value between the semi-mature and mature output signals (the greater of semi-mature or mature output signal becomes the cell context). The cell context is used to generate for each antigen an anomaly coefficient which will be used in the final step (classification) [32] to label the antigens as "normal" or "anomalous", (Figure 3.2, phase 3.) The context assessment phase can be seen as a 'local' classification at the DC scale, while the classification phase, which we will detail next, can be seen as a global classification at the DCs population scale.

3.4.1.4 Classification phase

After the context assessment phase, the anomaly coefficient (noted MCAV: molecular antigen value) reflects the degree of anomaly of a given antigen, and is defined as the proportion of mature DCs among the total number of DCs that have sampled the antigen (formula 3.2). Once the anomaly coefficient is calculated for each antigen, the algorithm can perform its classification task. This is done by comparing the anomaly coefficient of each antigen to an anomaly threshold. The anomaly threshold can be a user-defined parameter or can be generated automatically from the data.
The closer the anomaly coefficient is to 1, the greater the probability that the antigen is anomalous.

$$Anomaly \quad coefficient = \frac{\alpha}{\alpha + \beta} \tag{3.2}$$

where α and β are the number of mature and immature DCs that sampled the antigen, respectively (Figure 3.2, phase 4).

3.4.1.5 DCA a worked example

To illustrate the model, many studies [32, 16] have given a step by step application example of the DCA in the field of bank security.
In this context, the DCA is applied to the dataset presented in Table 3.1. Each client is seen as an antigen, with features such as the age, the number of credits, and income. The objective is to classify clients as "normal" or "anomalous" to decide whether to grant credit or not. The data item IDs (Table 3.1) represent the row numbers of the clients in this case.

Client	Age	Income	Number of credit cards	Duration of loan
ID 1	24	650	1	30
ID 2	30	1000	3	10
ID 3	36	1300	3	8

Table 3.1: An example of an input dataset where clients represent antigens (Ags), and the features: age, income, number of credit cards, and duration of loan' are used to generate the PAMPs, Ds, and Ss.

Pre-Processing and Categorization phase

DCA initially selects certain attributes and categorizes them into PAMPs, Ds, and Ss signal types. Expert knowledge is used to map these features to their most appropriate signal types. For example, PAMPs may include features like the number of credit cards, Ds might include the duration of loan, and Ss signals could include the age and incomes. The resulting dataset is then transformed into a signal dataset, as shown in Table 3.2. This transformation process is detailed in the subsequent section.

Client (Ag)	PAMP	Ds	Ss
Ag 1	100	0	100
Ag 2	0	100	0
Ag 3	20	50	40

Table 3.2: PAMPs, Ds, and Ss corresponding to the input dataset presented in Table 3.1.

Detection phase and Context assessment phase

To show the calculations under different input signals, three iterations (cycles) with three set of signals are shown. The derived output signal values are used to demonstrate how to perform the anomaly coefficient (MCAV) calculation for three antigens (Ag1, Ag2 and Ag3). Three DCs are required, one for each iteration, termed DC1, DC2 and DC3 for the purpose of identification.

The calculation of the output signals is given by Eq. 3.3, using the signal values shown in Table 3.2, and the weights presented in Table 3.3.

$$C[CSM, smDC, mDC] = (W_{PAMP} \times PAMP) + (W_{Ds} \times Ds) + (W_{Ss} \times Ss) \tag{3.3}$$

	W_{PAMP}	W_{Ds}	W_{Ss}
CSM	2	2	1
smDC	0	0	1
mDC	2	1	-2

Table 3.3: weights used for signals processing.

The worked example is performed as follows:

- Each antigen is randomly multiplied to form an antigen vector A:

  $$A = \{Ag1, Ag1, Ag1, Ag1, Ag1, Ag2, Ag2, Ag2, Ag2, Ag3, Ag3, Ag3\}$$

  where Nb-antigen (Ag1) $= 5$, Nb-antigen (Ag2) $= 4$ and Nb-antigen (Ag3) $= 3$.

- Cycle $l = 0$: DC1 randomly samples antigens from A, so

  $$DC1\, a(m) = \{Ag1, Ag1, Ag1, Ag2, Ag2\}$$

  where $a(m)$ is a sub-antigen vector and m is the DC index. DC1 samples input signals, so

  $$DC1\, s(m) = \{100, 0, 100\}$$

  where $s(m)$ is the signal vector of DC_m. DC1 calculates output signals using Eq. 3.3, so DC1 outputs:

  $$C_{CSM} = (100 \times 2) + (0 \times 2) + (100 \times 1) = 300$$
  $$C_{smDC} = (100 \times 0) + (0 \times 0) + (100 \times 1) = 100$$
  $$C_{mDC} = (100 \times 2) + (0 \times 1) + (100 \times -2) = 0$$

  DC1 has exceeded its migration threshold as the value for C_{CSM} is greater than $mt = 100$. Also, $C_{smDC} < C_{mDC}$ and therefore DC1 is assigned a cell context value of 1, indicating that its collected antigens may be anomalous.

- By removing the antigens already used by DC1, the antigen vector now consists of:

  $$A = \{Ag1, Ag1, Ag2, Ag2, Ag3, Ag3, Ag3\}$$

- Cycle $l = 1$: DC2 randomly samples antigens, so

  $$DC2\,a(m) = \{Ag2, Ag2, Ag1\}$$

  DC2 samples input signals, so

  $$DC2\,s(m) = \{0, 100, 0\}$$

  DC2 calculates output signals, so DC2 outputs:

  $$C_{CSM} = (0 \times 2) + (100 \times 2) + (0 \times 1) = 200$$
  $$C_{smDC} = (0 \times 0) + (100 \times 0) + (0 \times 1) = 0$$
  $$C_{mDC} = (0 \times 2) + (100 \times 1) + (0 \times -2) = 100$$

  DC2 has exceeded its migration threshold. Also, $C_{smDC} < C_{mDC}$, and therefore DC2 is assigned a cell context value of 1, indicating that its collected antigens may be anomalous.

- The antigen vector now consists of:

  $$A = \{Ag1, Ag3, Ag3, Ag3\}$$

- Cycle $l = 2$: DC3 samples antigens, so

  $$DC3\,a(m) = \{Ag1, Ag3, Ag3, Ag3\}$$

  DC3 samples input signals, so

  $$DC3\,s(m) = \{20, 50, 40\}$$

  DC3 calculates output signals, so DC3 outputs:

  $$C_{CSM} = (20 \times 2) + (50 \times 2) + (40 \times 1) = 180$$
  $$C_{smDC} = (20 \times 0) + (50 \times 0) + (40 \times 1) = 40$$
  $$C_{mDC} = (20 \times 2) + (50 \times 1) + (40 \times -2) = 10$$

  From the output calculation: DC3 has exceeded its migration threshold and the cell context value is 0 (immature DC)

Classification phase

- To generate the anomaly coefficients, we have to look for the antigens having a cell context equal to 1. Among the three DC cycles, we notice that in $l = 0$ and $l = 1$ the

cell context = 1. Therefore, to represent the Nb-mature variable, we have to count how many times each antigen is repeated in the DC1 $a(m)$ and DC2 $a(m)$ antigen vectors. For instance, in DC1 $a(m)$, Ag1 is repeated 3 times, thus,

Nb-mature (Ag1) $= 3$ and in DC2 $a(m)$, Ag1 is repeated 1 time, thus, Nb-mature (Ag1) $= 1$

Ag1 has been exposed 5 times in total, Hence,

$$\text{Anomaly coefficient (Ag1)} = \frac{4}{5} = 0.8$$

Similarly, Ag2 is repeated 2 times in DC1 and 2 times in DC2:

$$\text{Nb-mature (Ag2)} = 4$$

Hence,

$$\text{Anomaly coefficient (Ag2)} = \frac{4}{4} = 1.0$$

Ag3 does not appear in any context 1 vector:

$$\text{Nb-mature (Ag3)} = 0$$

Hence,

$$\text{Anomaly coefficient (Ag3)} = \frac{0}{3} = 0.0$$

- To perform antigen classification, a threshold (at) must be applied to the Anomaly coefficient. This threshold can either be generated automatically from data or be a user-defined parameter. Let us assume that $at = 0.4$. In this case, client1 (Ag1) and client2 (Ag2) are classified as anomalous which means that they are not allowed to have a credit. This is because their corresponding Anomaly coefficients are greater than the defined anomaly threshold. However, client3 (Ag3) is classified as normal (see Table 3.4).

Antigen type	Nb-antigen	Nb-mature	Anomaly coefficient
Ag1	5	4	0.8
Ag2	4	4	1.0
Ag3	3	0	0.0

Table 3.4: Anomaly coefficients of Antigens 1, 2, and 3.

3.4.2 DCA Improvements and Extended Versions

In this section, we provide an overview of the main improvements and extensions of the DCA.

Following the initial implementation of the DCA, [145] conducted a theoretical analysis that provided foundational insights, while [70] approached the DCA from a mathematical perspective, paving the way for subsequent advancements. One of the first improvements, implemented by the creator of the DCA, is the deterministic version known as dDCA [68]. In this version, the number of parameters has been reduced for simplification. A minimum of two signal categories is required: an activating signal with a positive weight (danger signals) and an inhibitory signal with a negative weight (safe signals), as shown in Eq. 4.1.

$$CSM = W_D \times S_D + W_S \times S_S \tag{3.4}$$

The pool of DCs is typically defined to be 100. In addition, the output context value of DCs is simplified to a single factor, in contrast to the first version where there were three output signals for each DC. Here, a positive value indicate a mature cell context and negative value indicates a semi-mature cell context.
Based on these improvements, many researchers have adapted and further developed the DCA to suit their specific needs. Most of them have proposed contributions in the pre-processing and categorization phase, addressing the signals and weight acquisition aspects. For example, [49] proposed an optimization technique to generate the set of optimal weights by using a genetic algorithm. [173] applied a numerical differentiation method in the pre-processing step to realize an adaptive acquisition of signals. [172] presented a model that combines a numerical differentiation for signal extraction with DCA that performs anomaly detection, then implemented it to earthquake prediction. [138] proposed a model employing Krill herd optimization for relevant feature selection, coupled with the DCA for identifying and classifying spam messages. [33] hypothesized that the DCA's sensitivity to the input class data order could be due to non-clear separation of contexts and noisy data, and thus proposed a hybrid fuzzy clustering approach to address this issue. [95] introduced a network intrusion detection algorithm based on the DCA, which incorporates multiresolution analysis and a segmentation approach to improve feature selection and signal categorization.

3.5 Conclusion

Literature review show that the DCA has distinct advantages when applied to real-time problems, as it does not require extensive training periods [172], and it has shown promising results by reducing high rates of false positives [111]. In addition, this method enables data items ('antigens') to be classified by integrating heterogeneous data through the use of two types of signals (danger and safe signals).
However, as seen in the previous examples, improvements have primarily focused on the pre-processing and categorisation phase. The DCA still relies on a large number of stochas-

tic elements and variable thresholds, which has drawn some criticism. Very few studies have addressed the migration threshold issue and random exposure in the detection phase, despite their crucial importance. As shown by [120], if a migration threshold is too low, a cell will migrate too quickly and will not be able to gather a representative sample of the input signals. If a migration threshold is too high, the cell will migrate too slowly and will misclassify the gathered antigens. Among the studies that have focused on this problem, [55] stands out as an example, focusing on building dynamism for the migration threshold of DCs and proposing a method for good sampling of antigens by DCs to ultimately generate a novel semi-supervised classifier.

Moreover, since its creation, the DCA has mainly been applied in the fields of computer security and anomaly detection. When it comes to epidemiological surveillance, one notable example is [111], who proposed an outbreak detection model based on danger theory.

In the next chapter (Chapter 4), we will explore how the DCA can be applied to the context of EBS (Event-based surveillance). To the best of our knowledge, this novel application has not been previously attempted. We will also address the limitations related to the migration threshold and the random exposure of the DCs, drawing inspiration from the studies cited previously to enhance the robustness and accuracy of the DCA in this new context.

Part II

Contributions

CHAPTER 4

TOWARDS A MODEL INTEGRATING EPIDEMIOLOGICAL AND ENVIRONMENTAL DATA FOR DISEASE SURVEILLANCE

In this chapter, we present the first contribution of the thesis: a model that integrates epidemiological and environmental data for event-based surveillance. This chapter introduces our proposed model, EpiDCA, which is based on the Dendritic Cell Algorithm (DCA) discussed previously in Chapter 3. We provide a comprehensive overview of the methodology underlying EpiDCA, detailing its four phases: the pre-processing and categorization phase, the detection phase, the context assessment phase, and the classification phase. Additionally, we apply our model to a first case study Avian Influenza (see Chapter 2, Section 2.2) and present the preliminary results of this initial implementation.

4.1 Introduction and objectives

In Chapter 3, we provided an overview of the danger theory and the DCA, which form the basis of our model. We presented the various works on DCA extensions and highlighted the limitations inherent to the method, which have been pointed out by several research studies. These limitations include the random exposure of the DCs in the detection phase, the setting of the migration threshold value in the context assessment phase, and the need to predefine the population of DCs before running the algorithm.
In this chapter, we present EpiDCA, an enhanced adaptation of the DCA specifically designed for application in the context of EBS. The objective of this model is to provide a robust and generic approach that can be easily applied to various epidemiological systems (zoonotic, vectorial, transboundary diseases), thus improving the EBS systems in terms of classification and early detection.
The specific contributions described in this chapter include:

- The development of a DCA based approach in an event-based surveillance context to combine epidemiological data and environmental data. To our knowledge, the DCA has not yet been implemented in EBS systems;
- The integration of spatial information in the Detection phase. Thus, the random detection of the DCA is replaced by a deterministic approach that depends on the spatial distance between the DCs and the antigens, extending the study described in [55];
- The integration of temporal information to determine the migration threshold;
- A detailed application and preliminary results from a first case study.

4.2 EpiDCA Workflow

In this section, we introduce the overall architecture of our proposed model, EpiDCA. Each subsequent subsection will provide a detailed description of a specific phase of the algorithm, highlighting the contributions of our method and presenting the corresponding pseudo-code for each phase.
In order to provide clarity, below are brief definitions of key concepts used in our context before diving into the details:

- **Antigens (Ags)**: In the context of our work, the events extracted from the detected articles by EBS systems represent our antigens (what we want to classify). Each event is characterized by the features: Date, Coordinates (location), affected Host, Source,

and Subtype. Initially, the anomaly coefficient of each incoming antigen is set to 0; it is calculated at the end of the process. See **Input** and **Output** in Algorithm 1.

- **Dendritic Cells (DCs)**: As in the original DCA, in our model, the DCs represent the instances that will transform the input signals (danger and safe signals) into an output signal (CSM).
  Instead of initializing a population of DCs, a new cell is created for each incoming antigen. See **Initialize** in Algorithm 1.

- **Danger signals (Ds)**: Epidemiological metadata extracted from detected events represent the danger signals. This metadata includes the source that reported the event (official or non-official), the name of the disease, the pathogen subtype (if available), and the host.

- **Safe signals (Ss)**: Refer to the environmental context or the environmental data extracted from disease suitability maps. The events' locations are extracted and associated by spatial correspondence with the corresponding environmental data.

Like the DCA, EpiDCA is divided into four main phases: the Pre-Processing and Categorization phase, the Detection phase, the Context Assessment phase, and the Classification phase (see Chapter 3 , Section 3.4.1), as illustrated in Figure 4.1.

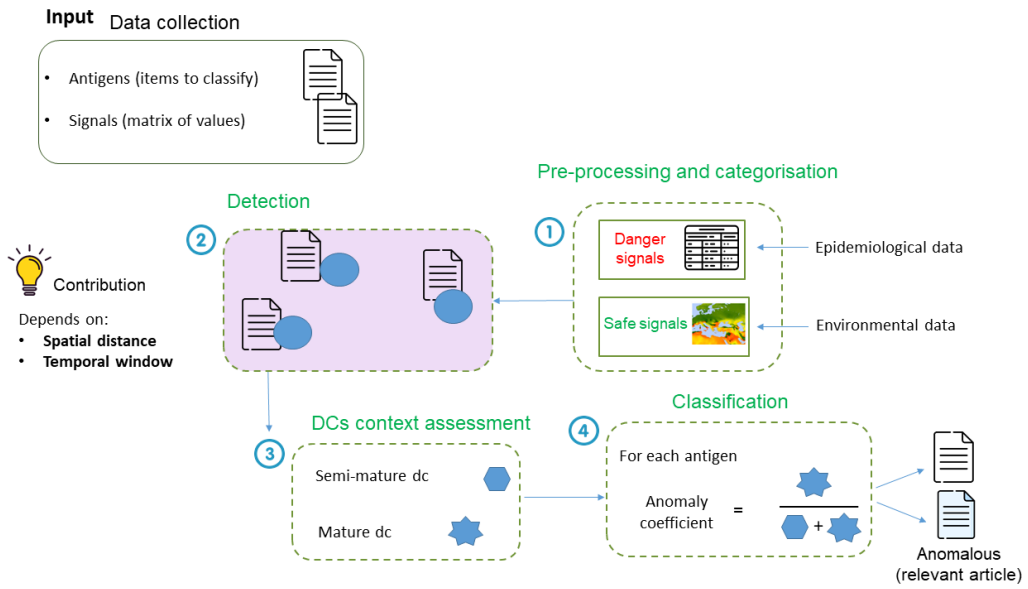


Figure 4.1: Towards a four-phase process for the event-based surveillance context.

The pseudo-code for the main EpiDCA phases is presented in **Algorithm** 1. Each phase will be explained in detail in the following subsections.

Algorithm 1 Main algorithm: EpiDCA

Input Antigens: set of $antigens\ (Date, Coord, Host, Source, Subtype, \emptyset)$, Disease
Output AGs (antigens classified as "normal" or "anomalous")
Initialize Empty set of cells: DCs
$AGs \leftarrow Antigens + ListCell = \emptyset + AnomalyCoef = Normal$
▷ ListCell is the list of cells that have been exposed to an ag - AnomalyCoef is set as Normal
▷ Call procedure to calculate Danger and Safe signals
1: $Pre\_Processing\_and\_Categorization\_phase(AGs)$
▷ Call procedure to detect DCs
2: $Detection\_phase(ag, DCs)$
▷ Call procedure to assess DCs context
3: $Context\_Assessment\_Phase(DCs)$
▷ Call procedure to classify antigens
4: $Classification\_Phase(AGs)$
▷ The anomaly coefficient is used to assess the degree of anomaly of a given antigen, the closer it is to 1, the greater the probability that the antigen is anomalous

4.2.1 Pre-Processing and Categorization phase

In the Pre-Processing and Categorization phase (Figure 4.1 Phase 1), input data (here, epidemiological data extracted from detected event and environmental data risks maps) are converted into two categories of signals: *danger signals* (Ds) and *safe signals* (Ss) as described in Chapter 3, Section 3.4.1.1. The epidemiological data extracted from the articles (source of information, host, disease) represent the danger signals. We refer to the knowledge of experts in order to establish a score and give a numerical value to each epidemiological data. For example, taking the example of AI: +30 if the source of information is official, +30 if the affected host is a wild bird, +40 if the highly pathogenic subtype is mentioned, etc (see Algorithm 1, Phase 1). Hence, if none of the AI epidemiological data are mentioned, the danger signal value is very low and the article is more likely to be discarded. These values were evaluated and refined empirically (see Section 4.3.5).
The environmental data represent safe signals. A maximum safe signal indicates that the environment is not favourable to the occurrence and the dissemination of the disease, a safe signal equal to 0 indicates on the contrary a suitable environment where risk factors are found.
Unlike the classic DCA versions where a group of DCs is initialized, in EpiDCA each antigen generates (or triggers the creation of) a new DC. Each DC is characterized by a date of creation, coordinates that specify its spatial location, an output signal (CSM), the Dc.Context (which can be mature or semi-mature), and a number of expositions (NbExp) that is used to calculate the context.
The complete pseudo-code to define the DCs creation is presented in **Algorithm 2: New Cell**.

Algorithm 2 New Cell

Input ag: $antigen\ (Date, Coord, Host, Source, Subtype, \emptyset)$, Disease
Output cell

1: $Ds \leftarrow DangerSignal\ (ag,\ Disease)$ (Eq1)
2: $Ss \leftarrow SafeSignal\ (ag,\ Disease)$ (Eq2)
3: $CSM \leftarrow ComputeSignal\ (Ds,\ Ss)$(Eq3)
4: **Return** $cell\ (Date,\ Coord,\ NbExp = 1,\ CSM,\ Context)$

Following the pre-processing and categorisation phase, *EpiDCA* is structured in three phases detailed in the following subsections.

4.2.2 Detection phase

In the detection phase, the DCs are exposed to the Ags (Figure 4.1 Phase 2). The *CSM_incoming* of the newly created cell is calculated by combining the weighted danger and safe signals as shown in Eq. 4.1.

$$CSM = W_D \times D_S + W_S \times S_S \tag{4.1}$$

Where W_D and W_S are the weights of Danger and Safe signals respectively.

The weighting can be adapted to the studied disease: For a disease strongly influenced by environmental factors (for example a vector-borne disease such as West Nile fever, or involving a wild reservoir, such as avian influenza) a greater weight could be given to the safe signals that define the environmental context in our approach.
In EpiDCA, the exposure of the DCs depends on: (1) the spatial distance between the DCs and the Ags with the radius of coverage R of the DCs and (2) a temporal window (the difference in days between the Ags publication date), (see Figure 4.2). At each time step, the cumulative output signals (CSM) of the DCs are updated as shown in Eq. 4.2:

$$\begin{cases} CSM_{t+1} = CSM_t + (\Delta_{dist} \times CSM_{incoming}) \\ CSM_0 = 0 \end{cases} \tag{4.2}$$

with Δ_{dist} a distance coefficient inversely proportional to the spatial distance, and is calculated as follow (Eq. 4.3):

$$\Delta_{dist} = \frac{\text{Disease.space\_limit} - \text{DistCellAnt}}{\text{Disease.space\_limit}} \tag{4.3}$$

Disease.space_limit represents the maximum distance within which two events can overlap. If an outbreak event occurs at a distance greater than this limit, the two events are con-

sidered unrelated and are not impacted by each other. This parameter depends on the disease being studied,
DistCellAnt represent the calculated distances between an incoming Ag (event) and the existing DCs (older events).
Equations 4.2 and 4.3 reflect that the greater the distance, the lower the contribution of the *CSM_incoming*, this translates the fact that the spread of certain diseases is linked to the distance between the observed events [135].

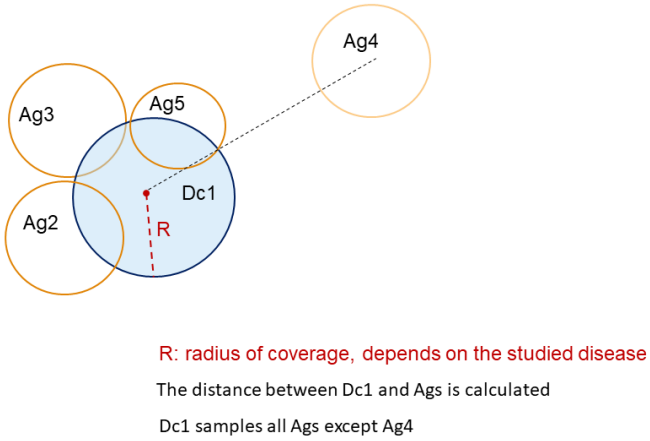


Figure 4.2: Example of a DC exposure.

Similarly, the temporal parameter *Disease.time_limit* refers to the maximum temporal distance within which two events can be considered related. If the time between events exceeds *Disease.time_limit*, the DCs will stop detection and proceed to the next step. This is why both spatial and temporal conditions are checked before calculating the CSM.
The complete pseudo-code for the detection phase is presented in **Phase 2 - Procedure Detection phase**.
The parameters *Disease.space_limit* and *Disease.time_limit* represent the spatial and temporal parameters used by the DCs, respectively. This aspect will be further explained in Subsection 4.3.2.

4.2.3 Context assessment phase

The context assessment phase takes into account the CSM and the number of exposures of each cell. At the end of this phase, each cell is labeled as "mature DC" or "semi-mature DC" (Figure 4.1, Phase 3). The "mature" label means that the cell concerned has been greatly exposed to danger signals during the defined period, unlike "semi-mature" cells. This information is then used in the Classification phase.
The CSM is defined through the calculation of the *Ratio_Exp*, which is a dynamic threshold

Phase 2 – 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**
  $DistCellAnt \leftarrow calculation\ distance\ between\ ag\ and\ dc$
  $Diffdays CellAnt \leftarrow calculation\ diffdays\ between\ ag\ and\ dc$
  if $DistCellAnt < Disease.space\_limit$ **and**
    $Diffdays CellAnt < Disease.time\_limit$ **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
$DCs \leftarrow DCs + NewCell(ag, Disease)$

that is calculated as the ratio of the mean CSM across all DCs to the number of DCs in the database, as shown in Eq 4.4.

$$Ratio\_Exp = \frac{Mean(DCs.CSM)}{Mean(DCs.NbExp)} \tag{4.4}$$

For each DC, if its CSM is greater than *Ratio_Exp*, it is considered mature. Otherwise, it is considered semi-mature.
The complete pseudo-code to define the cellular context is presented in **Phase 3 - Procedure Context Assessment phase**.

4.2.4 Classification phase

Finally, in the classification phase (Figure 4.1, Phase 4), the output signals are used to generate an anomaly coefficient specific to each antigen, which thus takes into account the epidemiological data extracted from the articles (danger signals), the environmental context (safe signals), and the spatio-temporal information of the events.
The Anomaly coefficient is calculated as the ratio of the total number of mature DCs to the

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

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

total number of DCs exposed to a given Ag, as shown in Eq. 4.5.

$$Coeff = \frac{ag.ListCell.mature()}{ag.ListCell.length()} \tag{4.5}$$

This anomaly coefficient value is between 0 and 1, the more it tends towards 1 the greater the probability that the antigen is anomalous. The anomaly threshold (AnomalyThreshold) is set at 0.5 as proposed in the literature [32]. The complete pseudo-code for the classification phase is presented in **Phase 4 - Procedure Classification phase**.

4.3 Methodology - First application on Avian Influenza

For reasons of data availability, we first implemented EpiDCA on the case study of AI in Asia. The following subsections describes the data collection, parameter settings, and anal-

yses conducted.

4.3.1 Data collection

AI is the focus of constant attention, and its events are reported in media articles by official sources (i.e., FAO, OIE, etc.) as well as non-official sources (i.e., online media, social networks, etc.) to promptly implement protection and control measures. We constructed a dataset from two EBS systems: HealthMap and PADI-Web [22]. Articles published between August 2018 and July 2019 were manually labeled as 'Relevant' or 'Irrelevant'. Relevant articles were those that reported at least one AI event (outbreak) (see example Figure 4.3). Articles labeled as irrelevant either described measures (e.g., economic, political, control measures) or were related to another disease. The dataset used for this first evaluation was initially collected for a previous work[158], and the articles were annotated by two epidemiologists (B. Boudoua and S. Valentin). We obtained two corpora:

- *DB_AI_Initial*: We first obtained an imbalanced corpus of 202 news articles (174 relevant and 28 irrelevant) related to AI only. This imbalance is typical in EBS contexts, where relevant articles frequently outnumber irrelevant ones. We first tested EpiDCA on this corpus. As a reminder, EpiDCA, is unsupervised, meaning no training phase is required. The purpose of the annotation is for classification evaluation, allowing us to compare the method's classification with the manual classification.

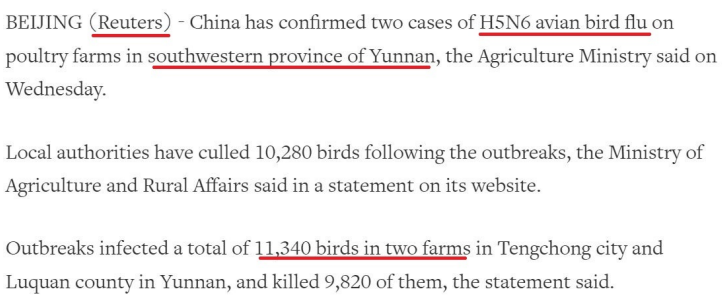
  To fairly compare our unsupervised method with state-of-the-art supervised machine learning techniques and to prevent bias towards the majority class, we extended this corpus to create:

- *DB_AI_Extended*: a balanced corpus of 348 articles (174 relevant and 174 irrelevant). In this case, the irrelevant articles either described economic/control measures or were related to African Swine Fever (ASF).

4.3.2 Parameters setting

Danger and Safe signals

Epidemiological metadata extracted from detected articles were used to generate danger signals (as presented in Table 4.1) consisting of three categories of parameters: the information



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.

Figure 4.3: Example of a relevant article detected by HealthMap, epidemiological metadata are underlined in red.

source (official/non-official), the host (e.g., wild birds, domestic poultry in the case of AI), and the subtype (e.g., highly pathogenic subtype vs low pathogenic subtype).

Articles ID	**Epidemiological data**			
	Source	**Subtype**	**Host**	
ID_1	FAO	HPAI	Wild birds	
ID_2	Twitter	Unspecified	Unspecified	
ID_3	OIE	LPAI	Domestic birds	
ID_4	Reuters	HPAI	Humans	
Antigens ID	**Danger signals**			
	Source	**Subtype**	**Host**	**Total Ds**
Ag_1	30	40	20	**90**
Ag_2	20	10	10	**40**
Ag_3	30	30	30	**90**
Ag_4	20	40	5	**65**

Table 4.1: Example of AI epidemiological metadata extracted from documents detected by HealthMap and PADI-web, and converted to Danger signals.

Each parameter category is defined by a minimum and maximum score, as outlined in Table 4.2. Initially, we established a range for each parameter category based on expert recommendations. Subsequently, experiments were conducted on a sample guided by these expert-recommended values to determine the most appropriate scores within these predefined ranges. Higher scores indicate greater relevance for epidemiological data. For example, official sources such as the OIE and the FAO are assigned higher scores than less reliable sources such as online media and social media, and Highly Pathogenic Avian Influenza (HPAI) scores higher than Low Pathogenic Avian Influenza (LPAI) and unspecified subtypes. It is important to note that the cumulative danger signal for each antigen must not exceed 100 according to the first DCA version established by [67].

To generate the safe signals, we created a suitability map for AI occurrence following the spatial multicriteria decision analysis approach developed by [143], updated with recent geographic datasets [23]. Risk factors used to build the suitability map included: domestic waterfowl density, chicken density, human population density, proximity to open water, proximity to areas suitable for rice-growing, and proximity to roads [143].

Parameter [min - max]	Label	Score
Source [20 - 30]	Official	30
	Non-official	20
Subtype [0 - 40]	HPAI	40
	LPAI	30
	Unspecified	10
	Else	0
Host [0 - 30]	Domestic birds	30
	Wild birds	20
	Unspecified birds	10
	Humans	5
	Else	0

Table 4.2: Overview table of the parameters and scores used to generate the Danger Signals.

In this map (Figure 4.4), suitability is expressed on a continuous scale ranging from 0 (lowest suitability) to 255 (highest suitability).

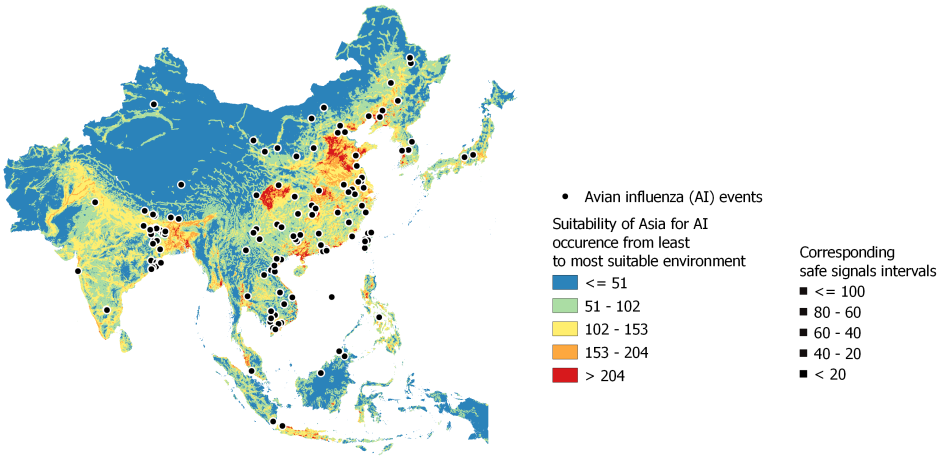


Figure 4.4: Suitability map of Asia for the occurrence of avian influenza (AI) in sensitive hosts, on a continuous scale from least to most suitable, along with the locations of AI events detected by EBS systems.

These suitability values were converted to safe signals applying a decreasing linear transformation (Figure 4.4). Safe signals thus lie within a range from 0 (the environment is suitable for AI occurrence) to 100 (the environment is not suitable for AI occurrence). Next, the events were associated with their environmental data by spatial correspondence using

QGIS[1], and the "point sampling tool" plugin[2] that allows one to assign to the events (points) the attributes (safe signal scores) of the underlying raster risk map (Figure 4.4).

DCs Coverage and migration threshold

We relied on expert knowledge to determine the radius of coverage and the migration threshold of the DCs. As AI viruses are likely to spread through different pathways (poultry transportation, wild birds migration, etc.) [170], it is difficult to precisely determine dissemination distance values. We set the DCs' radius of coverage at 20 km, a distance corresponding to the buffer zone where restrictions and control measures are implemented (surveillance zone) when AI outbreaks occur. Similarly, the migration threshold for DCs was set at 21 days [127, 9], because beyond this period, a location is considered free from AI if no new AI event is detected.

4.3.3 Classification analysis

This analysis aims to evaluate EpiDCA's capability to discern relevant events from irrelevant ones. From the following indices:

- True Positive (TP): Number of relevant articles correctly detected
- True Negative (TN): Number of irrelevant articles correctly detected
- False Positive (FP): Number of irrelevant articles incorrectly detected
- False Negative (FN): Number of relevant articles incorrectly detected

We calculated these metrics: precision, recall, and F-score, which are defined as follows:

$$Precision = \frac{TP}{TP + FP} \tag{4.6}$$

$$Recall = \frac{TP}{TP + FN} \tag{4.7}$$

$$\text{F-score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} \tag{4.8}$$

These metrics were calculated for each class (relevant and irrelevant), the results and the weighted average F-scores are presented in Section 4.4.

[1] https://www.qgis.org/de/site/
[2] https://plugins.qgis.org/plugins/pointsamplingtool/

Our method was tested on the imbalanced corpus *DB_AI_Initial*. A first test was carried out with fixed Ds values, meaning that all the epidemiological data used to compute Ds had a unique score, and without taking into account the Ss. Then, a second test with Ds computed with the scoring defined by the experts (Table 4.2), and without taking into account the Safe signals (Ss). Finally, a third test with both Ds and Ss was carried out.

4.3.4 Reactivity analysis

In our context, as discussed in Sections 3 and 4, each incoming antigen triggers the creation of a new DC that is initially immature. Based on the signals it detects, this DC differentiates into a mature cell (indicating an anomaly) or remains semi-mature.
Reactivity is defined as the time difference, measured in days, between the maturation date of the DCs and the confirmation date of an event that occurred at the same location. This measure of reactivity allows us to assess whether our system is capable of early detection of outbreaks. The reactivity of each mature DC is calculated as:

$$\text{Reactivity} = \text{Maturation Date} - \text{Confirmation Date} \tag{4.9}$$

The conditions can be summarized as follows:

- Reactivity > 0: Late detection. This suggests that the detection happened later than the confirmation date.
- Reactivity $= 0$: Detection occurred on the same day as the confirmation.
- Reactivity < 0: Early detection. This suggests that the detection happened earlier than the confirmation date.

The reactivity analysis was conducted on AI-confirmed events. We relied on the AI outbreaks data reported by the IBS system EMPRES-i, which allows us to download structured data regarding AI outbreaks in both humans and birds (domestic and wild). Epidemiological data such as the subtype, host, number of cases, number of animals culled, and the location at the administrative, regional, and district levels are provided.
Knowing that each article is characterized by its publication date and each DC is characterized by its maturation date, the calculation of the reactivity index involves the following steps:

1. **Linking Events:** Each event in the *DB_AI_Initial* database that has been confirmed is linked to its corresponding confirmation date (the date of official notification) from the EMPRES-i database.

For this corpus, the linkage of detected events with confirmed events is facilitated by the detailed information provided in both articles and the EMPRES-i database. Both sources include epidemiological data such as the event subtype, host species, location, and the number of affected and culled individuals. This level of detail, which is also common in online media reports, allows for precise matching. See Figure 4.5 for an example of an AI article detected by HealthMap, and Table 4.3 for how the event is reported in the EMPRES-i database, which is considered as a gold-standard because it contains officially confirmed events.

2. **Reactivity Calculation:** The reactivity of each mature DC is calculated using Eq. 4.9 . This allows us to assess the timeliness of our detection system in relation to the confirmed events.

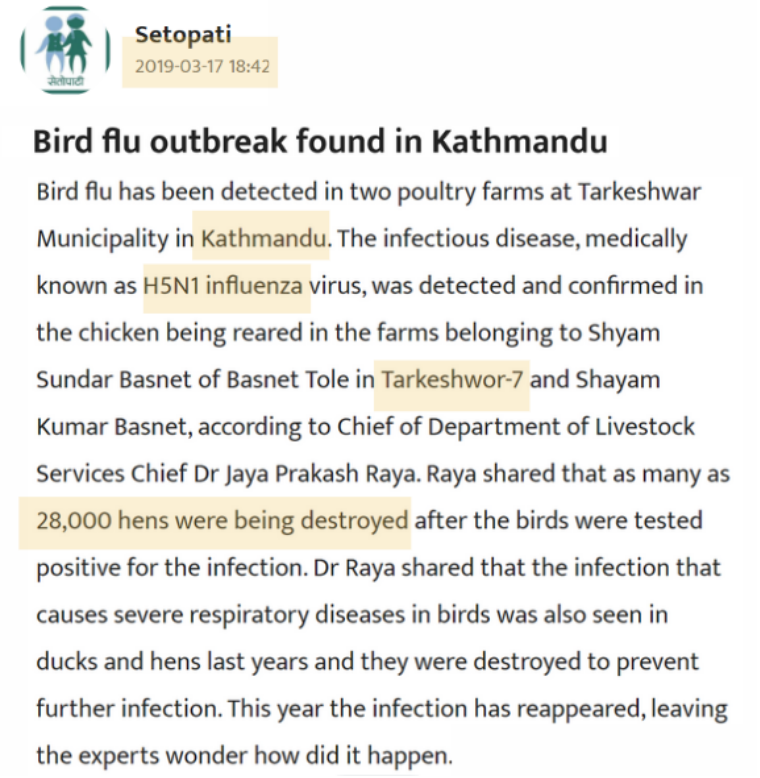


Figure 4.5: Example of an article detected by HealthMap.

Id	latitude	longitude	country	locality	reportingDate	subtypes	species	sumCases	sumDestroyed
249487	27.7714	85.3186	Nepal	Tarkeshwor	23-03-24	H5N1	domestic, chicken	2125	28548

Table 4.3: Extract of metadata associated with the AI event reported in Figure 4.5 stored in the EMPRES-i database.

4.3.5 Sensitivity analysis

The sensitivity analysis we conducted was divided into two phases. In the initial phase, we aimed to systematically evaluate the scores assigned by experts to various parameters. The second phase used the Morris method One At a Time (OAT) to identify parameters that had a more significant influence.

Phase 1: Sensitivity to the scores assigned by the experts

The first phase of the sensitivity analysis was conducted using the method described in [124]. For each parameter (Subtype/Disease Name, Host, and Source), the range of scores was defined by the minimum and the maximum values given by different experts Table 4.2. The sensitivity analysis was conducted according to the following steps:

1. We choose the score value we want to test.
2. We use the 4.10 formula to adjust the other scores accordingly:

$$Si = \frac{(100 - Sm) \cdot Si0}{(100 - Sm0)} \tag{4.10}$$

   where $Sm0$ and $Si0$ are the initial scores of the main changing parameter and the i-th parameter in the base model, respectively. This formula allows adjusting the other scores while ensuring that the total Ds score does not exceed 100.

3. We conducted a series of simulations to assess the impact of the new scores distribution. Each simulation involved multiple iterations where we systematically modified one or two parameters at a time. We adhered to the hierarchical scoring system provided by experts.
4. Additionally, we performed one iteration in which each group of parameters (Source, Subtype/Disease name, and Host) was set to zero. This approach allowed us to assess the individual impact of each parameter group on the overall model performance.
5. After each run, we calculated precision, recall, and F-score for both the negative and positive classes.

A total of 100 iterations was conducted on $DB\_AI\_Initial$. All the results of these simulations are available in the appendix.
The objective of this approach is to evaluate the effects of the new scores across various scenarios, and to understand the individual impact of different parameter groups on model

performance.

Phase 2: Morris method (One step at a time method)

The Morris method [142], is a one-step-at-a-time (OAT) global sensitivity analysis technique, where only one input parameter is adjusted per run. This method involves the following steps:

1. **Elementary Effects Calculation:** Local sensitivity measures, known as *elementary effects*, are calculated by measuring the perturbation in the output of the model when one parameter is changed.

2. **Statistical Analysis:** From the distribution of elementary effects, two key statistics are derived:

   - **Mean (μ^*):** The mean of the absolute values of the elementary effects. A high mean indicates a parameter with a significant impact on the output.
   - **Variance (σ):** The variance of the elementary effects. A large variance suggests that the parameter either interacts with other factors or has a non-linear effect on the output.

Usually, these thresholds are used for the OAT sensitivity analysis:

- Parameters with negligible effects ($\mu^* < 0.1$),
- Parameters with linear effects on the output and without interaction between parameters ($\sigma < 0.1$),
- Parameters with interactions and/or nonlinear relationships ($\mu^* > 0.1$ and $\sigma > 0.1$).

The objective of the OAT method is to evaluate each parameter individually and identify those with significant influence.
In the following section, we present and discuss the classification results, reactivity results, and sensitivity analysis results

4.4 Results and discussion

Classification results

This section presents the classification results of EpiDCA on both corpora (*DB_AI_Initial* and *DB_AI_Extended*) introduced in Section 4.3.1
The results presented in Table 4.4 with the imbalanced corpus show a high precision value and a lower recall value for the relevant class, while the opposite pattern is observed for the irrelevant class. The best results were obtained when both Ds and Ss were computed. These results suggest that an appropriate Ds scoring, based on epidemiological data, is crucial for effective classification. Furthermore, considering the environmental context, particularly through its spatio-temporal dimensions via the safe signal, significantly improves the accuracy of identifying and classifying articles detected by EBS systems.

Fixed Ds values	Precision	Recall	F-score
Relevant class	0.861	0.534	0.659
Irrelevant class	0.138	0.464	0.213
Weighted average	**0.759**	**0.524**	**0.529**
Without Ss			
Relevant class	0.980	0.569	0.720
Irrelevant class	0.257	0.928	0.403
Weighted average	**0.879**	**0.619**	**0.675**
Computed Ss and Ds			
Relevant class	0.971	0.787	0.869
Irrelevant class	0.393	0.857	0.539
Weighted average	**0.889**	**0.797**	**0.823**

Table 4.4: EpiDCA classification results on the *DB_AI_Initial* corpus. For each test, the evaluation metrics (precision, recall and F-score) were calculated per class. The last row indicates the weighted average scores.

Then, additional tests were conducted on the balanced corpus (*DB_AI_Extended*), in order to fairly compare EpiDCA with supervised machine learning methods. We used Weka[3] software to test four supervised learning methods (SVM (with a polynomial kernel), Naive Bayes, K-nn and Random Forest) by performing a 5-fold cross-validation. We obtained F-scores comprised between 0.868 (Naive Bayes) and 0.908 (Random Forest). The classification results are presented in Table 4.5 for EpiDCA and in Table 4.6 for the supervised learning methods.

The results confirm that an appropriate danger signals categorization and scoring in the

[3]https://www.cs.waikato.ac.nz/ml/weka/index.html

pre-processing phase is crucial for a good classification. Very good results (precision, recall and F-score greater than 0.9) were obtained with computed Ds, with or without Ss. Of note, even better results were obtained without including the Ss (Table 4.5), which is probably an artefact due to the irrelevant ASF events included in the *DB_AI_Extended* to obtain a balanced dataset. Indeed, ASF is a disease that differs from AI in terms of hosts and risks factors [126]. For this reason, it is likely that the use of the same risk map (safe signals) and parameters to classify these events biases the evaluation.

Fixed Ds values	Precision	Recall	F-score
Relevant class	0.514	0.511	0.513
Irrelevant class	0.514	0.517	0.515
Macro average	**0.514**	**0.514**	**0.514**
Without Ss			
Relevant class	0.885	0.936	0.910
Irrelevant class	0.932	0.879	0.905
Macro average	**0.908**	**0.907**	**0.907**
Computed Ss and Ds			
Relevant class	0.897	0.908	0.902
Irrelevant class	0.907	0.896	0.901
Macro average	**0.902**	**0.902**	**0.901**

Table 4.5: EpiDCA classification results on the *DB_AI_Extended* corpus. For each test, the evaluation metrics (precision, recall and F-score) were calculated per class. The last row indicates the macro average scores.

Unlike the supervised methods, EpiDCA produced consistent detection results and effectively differentiated between relevant and irrelevant articles, all without requiring a training phase. Moreover, since no training phase is required, the proposed approach is able to process real-time input and has significant potential for implementation as a real-time system. The effectiveness of EpiDCA can be attributed to the use of expert-defined parameters and its integration of spatio-temporal information. Both of these aspects are crucial in the effective analysis of epidemiological data.

Another key feature of the proposed method is the ability to adjust the scores given to the epidemiological data in order to generate the danger signals. This offers the possibility to highlight a specific type of AI events (wild birds cases, domestic birds cases, or human cases) simply by modifying the danger signals parameters.

Method	Precision	Recall	F-score
SVM	0.895	0.894	0.894
Naive Bayes	0.869	0.868	0.868
K-nn	0.893	0.891	0.891
Random Forest	**0.916**	**0.908**	**0.908**

Table 4.6: Classification results with supervised machine learning methods on the *DB_AI_Extended* corpus.

Reactivity analysis results

Results demonstrate the effectiveness of our method in early detecting events. Out of the 202 DCs analyzed, 62.38 % (126/202) were classified as matured, with reactivity ranging from 36 days before to 29 days after official confirmation. Impressively, 34.05% (43/126) of the matured DCs were associated with early detected events. Notably, early detected events were primarily associated with AI cases in wild birds. In contrast, human cases were all confirmed on the same day or after official confirmation. For domestic birds, the detection patterns were more varied, with events detected both before and after confirmation, as shown in Figure 4.6.

This variation can be explained by the structured monitoring systems in place for domestic birds and humans, such as farms surveillance and medical reporting, which in most cases ensure timely detection and reporting. In contrast, the surveillance of wild birds is challenging due to difficulties in obtaining samples, often relying on targeted populations over localized areas or the reporting of opportunistically found dead animals or live birds caught for other reasons [139]. These challenges align with studies on AI, which indicate that EBS systems are particularly effective in the early detection of AI in wild birds [158, 59].

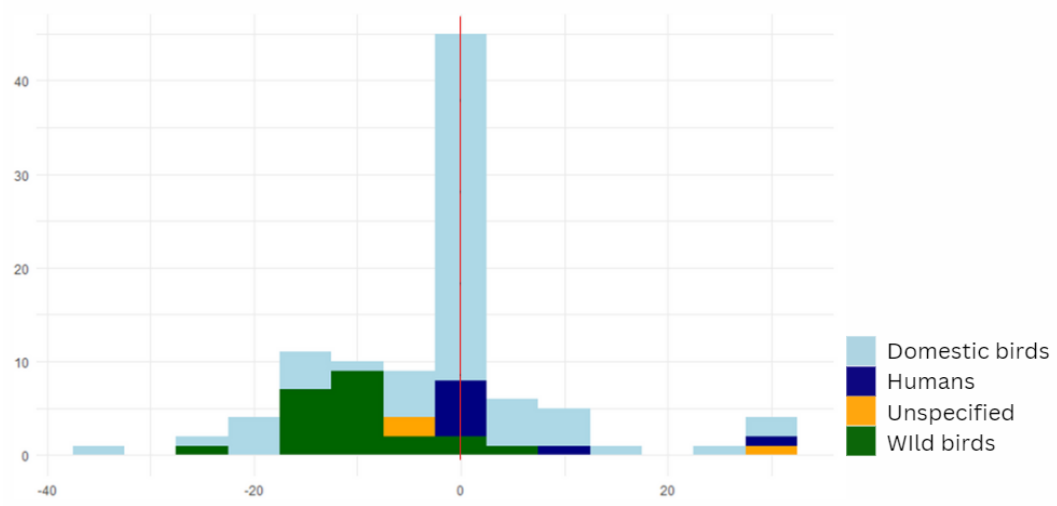


Figure 4.6: Reactivity of EpiDCA to AI.

Sensitivity analysis results

Overall, the model exhibited robustness, with minimal changes in results. The weights assigned by the experts appeared to be the most suitable for both corpora.
In our analysis using the Morris method (OAT), we observed that all parameters exhibited a negligible effect on the results, as indicated by their σ^* values being less than 0.1, as shown in Figure 4.7. Results indicate that the source parameter had no impact on the classification

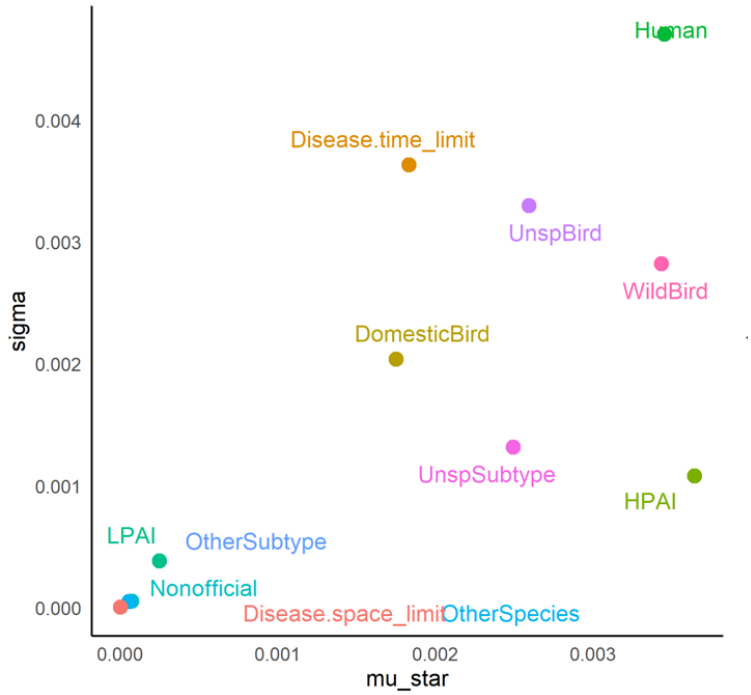


Figure 4.7: Morris OAT results for *AI_Initial*. The graph represents the average of elementary effects in absolute values (μ^*) according to their standard deviation (σ) with respect to model outputs.

results. This lack of impact is due to the fact that both corpora used exclusively consisted of non-official sources.
The Subtype parameter group proved to have the most significant impact. When setting this parameter to 0, the F-score dropped from 0.8 to 0.6. Specifically, the HPAI parameter had a notable impact on the classification; this can be attributed to the precision associated with specifying the subtype, indicating an event that is already confirmed at a local scale.
The Host parameter group demonstrated a significant influence, and assigning a value of 0 to the 'other host' label proved crucial for filtering out irrelevant events.
The temporal parameter appeared to have an important impact on the classification. However, assessing the temporal and spatial parameters separately poses challenges due to their inherent interdependence.
When the time window is too short, very few events occur within a specific zone. Consequently, even if we adjust the spatial parameter to different distances, the resulting changes

may not be discernible. This is why the outcomes of the OAT method indicate that the spatial parameter is negligible. Further experiments were conducted by adjusting this parameter to greater distances (over 1500 km) and extending the temporal parameter. In these instances, DCs systematically sample all incoming antigens, transitioning from immature to mature states. As a result, the anomaly threshold (the ratio of mature to immature DCs) consistently exceeds 0.5, leading to the classification of all antigens as abnormal. It is important to note that the integration of the spatial parameter in EpiDCA is a key contribution, as it transforms the previously random exposure of DCs to antigens into a deterministic process.

4.5 Conclusion

In this chapter, we presented EpiDCA along with our methodological contributions, including a first case study application on avian influenza in Asia. We evaluated EpiDCA in terms of classification performance, reactivity and sensitivity to parameters. In terms of classification, EpiDCA successfully detected relevant events in both unbalanced and balanced corpora (*DB_AI_Initial* and *DB_AI_Extended*, respectively). Very promising results, comparable to those of well-known supervised methods, were obtained.
Regarding reactivity, EpiDCA effectively identified officially confirmed events, with particularly strong results observed in the context of wild birds. Additionally, the model demonstrated robustness, indicating that the chosen parameters are well-suited for its application on AI cas study.
Our next objective is to test the robustness and genericity of EpiDCA when applied to different case studies and geographical contexts.
In the next chapter (Chapter 5), we will introduce a method and guideline for annotating articles according to their epidemiological topic. This involves three types of diseases: a zoonotic disease (Avian Influenza - AI), a cross-border animal disease (African Swine Fever - ASF), and a vector-borne disease (West Nile Disease - WND). The resulting dataset will be used to evaluate EpiDCA.

CHAPTER 5

AN ANNOTATION METHOD AND AN ORIGINAL DATASET FOR EVENT-BASED SURVEILLANCE OF AI, ASF AND WND

In the previous chapter, we introduced EpiDCA and demonstrated its initial application on the case study of AI in Asia, using two corpora: *DB_AI_Initial* and *DB_AI_Extended*. In this chapter, we present an original annotation method and guidelines that we used to create an annotated dataset. This dataset is designed to evaluate EpiDCA across different geographical contexts and diseases, including Avian Influenza in France, African Swine Fever, and West Nile Disease in Europe.

5.1 Introduction and Objectives

This chapter aims to present an annotation method and guidelines designed to produce an original annotated dataset. The annotations regarding epidemiological topics, combined with the extraction of epidemiological metadata and location information, will enable us to evaluate the robustness and genericity of EpiDCA on different geographical contexts and diseases, including Avian Influenza in France, African Swine Fever, and West Nile Disease in Europe. EBS systems usually classify articles as relevant or irrelevant by relying on human moderation or by implementing classification algorithms. These systems use annotated data to improve their classification in terms of accuracy and thus swiftly detect outbreak events. Consequently, the performance of these algorithms is highly dependent on the quality of the dataset used to train them [114].

As previously discussed in Chapter 1, the classification of epidemiological texts and the information extraction poses several challenges, as noted by [161, 157]. These challenges include the ambiguity often found in epidemiology-related texts, where disease mentions may not necessarily indicate an outbreak but could instead provide general disease information or historical context in a specific area. Furthermore, several locations can be mentioned within the text and at different levels of granularity [147], and the notion of 'Relevance' doesn't have a formal definition [161].

In this chapter, we present a method and a guideline that allow annotating articles according to their epidemiological topic when dealing with three different types of diseases: a zoonotic disease (Avian Influenza - AI), a cross-border animal disease (African Swine fever - ASF) and a vector-borne disease (West Nile Disease - WND). Metadata, such as the disease name, events' location, host, and virus subtype (specifically for AI), were manually extracted with special attention given to the outbreaks' locations reported in articles.

The contribution described in this chapter can be divided into two parts. First, it offers a detailed and reproducible annotation method that enhances the precision and reliability of epidemiological datasets. The annotation method and guidelines we present are designed to be generic, and can also be used to annotate datasets for the same diseases or serve as templates for annotating datasets related to diseases that share common characteristics: the WND guidelines can be used as a template for other vector-borne diseases, the AI guidelines can be used for other zoonotic diseases, and the ASF guidelines can be used for other transboundary diseases.

Second, the dataset produced will be used to validate EpiDCA but can also serve as a valuable resource for training supervised machine learning methods and fine-tuning language models.

5.2 Methodology

5.2.1 Construction of the dataset

We collected articles that were published within recent periods characterized by occurrences of outbreak events. Specifically: For AI, we focused on articles published online between August 2022 and January 2023 in France, as this period was marked by a significant number of AI events in the country [10]. Regarding ASF, we selected articles published between April and July 2022 in Europe. This period was chosen as it was characterized by new introductions and occurrences of ASF in some European countries [47]. For WND, we focused on articles published between June and September 2022 in Europe. As a vector-borne disease, the number of recorded WND cases is higher at the end of summer due to the increasing activity of its transmitting vectors (mosquitoes) during this period [46].
Extracted articles were provided with the following information: ID, title, source, publication date and URL. For each disease, a corpus composed of manually annotated news articles was compiled. Due to copyright reasons the texts of articles are not stored in the database but remain accessible via the provided URLs. The production of these corpora revolved around two axes: first, the articles annotation by relevance, and second, the epidemiological metadata extraction. The label's definition was built upon the framework described by [159]. The annotation was done at the document scale, with the annotator assigning one of three primary labels to each article: Relevant event, Relevant general information, or Irrelevant. While each category includes different subclasses with more nuanced definitions, we simplified the annotation process by using these three main categories. The definitions of these categories are as follows:

- **Relevant (events)**: Articles that clearly describe at least one epidemiological event along with its location. This category includes three subclasses: Confirmed cases, Persistent outbreaks, and Warning signals. Example: "The highly pathogenic bird flu, avian influenza, has been detected last week in a duck farm in France's eastern department of Ain, causing a total of 10,600 ducks culled." [166];
- **Relevant (general information)**: Articles that do not directly refer to a specific outbreak but still provide information on the incidence of a disease in the area studied. This category includes four subclasses: Assessment of the number of outbreaks, General consequences, Absence of new cases since a specific time, or Highlights of old events only. For example: "A novel research from the University of Extremadura (UEx), published in the journal Veterinary Microbiology, demonstrates the circulation of West Nile Virus in small wild birds (passerines) within a radius of about 15 kilometers from the city of Badajoz." [48];
- **Irrelevant**: Articles that mention the disease name without necessarily containing

relevant epidemiological information on the incidence of the disease in a given region. This category includes either preparedness articles or articles that are too vague in terms of spatial information. For example: "Swine fever: Dead deer: New fence in the national park is ready. With the relocation of the fixed fence in the protection corridor along the border with Poland over a length of eleven kilometers to the west, wild animals now have enough space to withdraw from the floodplain areas regularly affected by flooding." [8]

If both "event" and "general information" types were found in the same document, the "event" label takes priority. In cases where multiple outbreak events were reported within a single document, the article was duplicated for each mentioned event. Thus, the resulting number of articles (717 articles for AI, 300 for ASF, and 409 for WND) include duplicates; events with the same ID were extracted from the same article. The detailed guidelines, along with the manually annotated corpora, are available in an open source data repository [24].
In addition to the annotation by relevance, the epidemiological metadata associated with each article were manually extracted. It includes: the publication date, the disease name, the host, and the virus subtype (specifically for AI). A particular emphasis was given to the event's spatial information. In epidemiological texts, the location is provided at different levels of granularity. At a minimum, the continent/country is mentioned, and the department and/or the city can also be specified. For each article, we manually extracted all the cited locations and then associated the most detailed granularity location with its latitude and longitude coordinates. Example: "The highly pathogenic bird flu, avian influenza, has been detected in a duck farm in France's eastern department of Ain, causing a total of 10,600 ducks culled." In this example, the AI event is associated with the coordinates of Ain, France. Table 5.1 provides a summary of the different categories with examples for each subclass.

5.2.2 Guidelines design

The guidelines were designed according to an iterative process made of two annotation rounds, both performed by two epidemiologists (B. Boudoua and M. Richard). The objective was to make the guideline as generic as possible and as precise as necessary so that non-expert annotators could annotate these diseases related articles without running into ambiguity issues.

First, a preliminary guideline version was established with concise labels definitions that aim to describe the main content of the articles (Figure 5.1, Step 1) as 'Relevant' - this class includes articles that mention at least one outbreak event; or 'Irrelevant' - this class includes articles that don't contain any epidemiological information. During the first annotation round and for each case study, the experts were asked to annotate independently and

Table 5.1: Guideline definitions and examples.

Category	Subclass	Class Description	Example
Relevant Event	Confirmed cases	Correspond to animals/patients who have been tested positive for the disease.	“A new case of West-Nile virus has been detected in Gonimo Serres, in an 87-year-old woman, according to Serres medical examiner.”
	Persistent outbreak	Correspond to outbreaks that last over time in a given region. New cases are detected regularly to confirm that the disease is still circulating (at the time of publication of the article). It can also correspond to articles containing an aggregation of events between a prior date and the current date. Keywords such as "since", “so far” or "until" suggest that the disease is still present.	“Until 06/09/2022, cases of West Nile virus infection have been recorded in Greece, in settlements in the Metropolitan Unit of Thessaloniki. It is considered possible and expected to diagnose further cases in the coming period.”
	Warning signals	It can correspond to several types of information depending on the disease: Suspected cases (dead or sick animals with high suspicion of disease but without confirmation by diagnostic tests) and imported cases (the origin of the outbreak is considered as a warning signal) for the three diseases. For ASF, it also pertains to viruses detected in animal products as well as cases reported at borders. For WNV it concerns viruses detected in mosquitoes.	“Food and Safety Agency is carrying out regular surveillance after dead birds were found in several locations.”
Relevant General information	Assessment of the number of outbreaks	Articles that don’t refer to a particular outbreak but give an overall picture of the number of outbreaks in a place or lists the most regions affected by the disease, i.e. it indicates the trend or the endemic status of the disease in an area.	“In Spain, West Nile virus can already be considered endemic in Western Andalusia and Catalonia.”
	General consequences	No specific outbreak is mentioned but it’s easy to understand that the disease is present in the reported area by the consequences it causes (economic, commercial, political or sanitary).	“The compensatory allowances granted to farmers in Brandenburg for damage caused by African swine fever (ASF) have almost quintupled in one year.”
	Absence of new cases since a specific time	It can correspond to the absence of new cases since the last case identified or since the lifting of restrictions.	“Belgium has been declared free since December 21, 2021. That is, African swine fever has been eradicated from the territory, no new cases have appeared.”
	Highlighting old events only	Articles that only mention past outbreaks or an aggregation of events between two past dates. The article cites the outbreaks not to warn of a new outbreak but rather to provide general information.	“The last detection of West Nile virus of the season occurred on December 12, 2021 in an equine in Germany.”
Irrelevant	Preparedness	The news refers to preventive measures without mentioning any cases: it can be the alert status.	“Although no cases of ASF have been detected, South Korea has introduced new measures to prevent the introduction of the disease.”
	Irrelevant or vague spatial information	Articles that refer to cases but in a different continent/country that the one we study or with a too large scale. For example, if we are working on the scale of Europe, mentioning only the North/South/East/West of Europe is not specific enough to provide relevant information (at a minimum, the country must be specified).	“There are still many infections in Eastern and South Eastern Europe, both in wild boars and domestic pigs.”

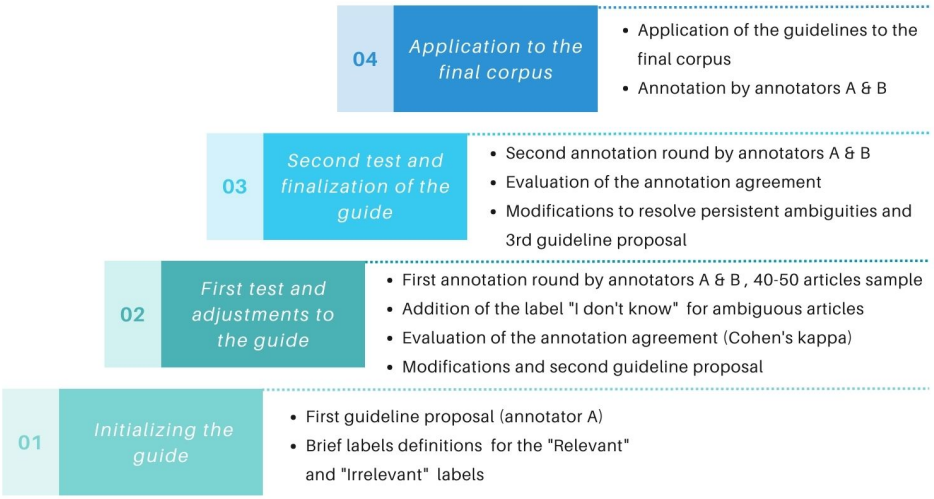


Figure 5.1: Pipeline of the annotation guideline elaboration process.

blindly a sample (n) made of 40 to 50 articles extracted from PADI-web. The annotators had to choose one single label for each article based on its relevance for epidemic intelligence purposes. Subsequently, the annotation agreement was calculated using the Cohen's Kappa coefficient [109]. Then, the annotation disagreements were discussed among the experts and this process led to the reformulation and refinement of the labels definition (Figure 5.1, Step 2).

Once these modifications were integrated, a second annotation round was done by the experts. To ensure the clarity and precision of the guidelines, we calculated the annotation agreement after each annotation round, using Cohen's Kappa coefficient. This coefficient measures the agreement between two observers during qualitative coding into categories, and is calculated as presented in Eq. 5.1:

$$K = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \tag{5.1}$$

where Pr(a) is the proportion of agreement between annotators and Pr(e) the probability of random agreement. The kappa values are between –1 and +1 and the higher the value, the stronger the agreement (K=1: perfect agreement). (Figure 5.1, Step 3). Following the integration of all the modifications, the final guideline version described in Table 1 was used to annotate the corpora (Figure 5.1. Step 4).

5.3 Results and discussion

Two rounds of annotation were required for both AI and ASF corpora before obtaining a satisfying Cohen's Kappa coefficient. A very good result was obtained during the initial annotation round for the WND corpus. This outcome underscores the effectiveness of our approach in designing the annotation guidelines. Our experience with the first two corpora (AI and ASF) enabled us to refine and develop guidelines that are both relevant and generic. These refined guidelines were easily applied to the new corpus (WND).
Table 5.2 compares the agreement results obtained in the first annotation round (initial version of the guidelines), and the second annotation round (final version of the guidelines). Results obtained for both AI and ASF during the initial round of annotation indicate a moderate agreement, with Cohen's kappa coefficients of 0.43 and 0.61 respectively. In contrast, the results obtained after the revision of the guidelines reflect a significant to almost perfect agreement with Cohen's kappa coefficients of 0.78 and 1 for AI and ASF respectively. For the WND corpus, a very good result of 0.84 was obtained during the first round of annotation, highlighting the relevance and genericity of our guidelines.

	AI (n=48)	ASF (n=40)	WND (n=50)
Annotation Round 1	0.427	0.605	**0.844**
Annotation Round 2	**0.78**	**1**	

Table 5.2: Cohen's Kappa values obtained during the two rounds of annotation.

Following annotation, we obtained three corpora:

- *DB_AI_France*: a corpus of 717 articles published between August 2022 and January 2023
- *DB_ASF_Europe*: a corpus of 300 articles published between April and July 2022
- *DB_WND_Europe*: a corpus of 409 articles published between June and September 2022

This allowed us to highlight specific characteristics for each case study. Table 5.3 illustrates the differences observed in label distribution, clarity of articles, and spatial information precision among events of AI, ASF, and WND.

5.4 Conclusion

In this chapter, we presented an annotation method and guidelines that can be used as a reference for annotating datasets related to the same diseases described here or serve as templates for annotating datasets related to diseases with similar characteristics. For example,

	DB_AI_France	*DB_ASF_Europe*	*DB_WND_Europe*
Observations noted during annotation	Ambiguous articles (containing several types of information, several likely labels)	Mostly clear and unambiguous articles (containing only one type of information)	Ambiguous articles (containing several types of information, several likely labels)
Relevance Annotation Observations	Majority of articles with the "event" label (**63%**)	Majority of articles with the "event" label (**79%**) including a lot of "Warning signals" (**11.3%**)	Majority of articles with the "event" label (**70.2%**)
	21.1% of articles with the "general information" label	**11.6%** of articles with the "general information" label	**10.3%** of articles with the "general information" label
	15.9% of irrelevant articles	**9.3%** of irrelevant articles	**19.6%** of irrelevant articles
Spatial Information	4 levels of precision for the mentioned scale: country, region, department and city	3 levels of precision for the mentioned scale: country, region and city	3 levels of precision for the mentioned scale: country, region, and city
	The city is mostly mentioned (**62.6%**)	The city is mostly mentioned (**43%**)	The city is mostly mentioned (**63.3%**)

Table 5.3: Specificity and differences observed during the annotation for each disease.

the WND guidelines can be used as a template for other vector-borne diseases, the AI guidelines for other zoonotic diseases, and the ASF guidelines for other transboundary diseases. The resulting dataset can be used for training supervised learning methods and/or fine-tuning language models. Additionally, the locations extracted from the document allow for testing spatial-based methods across different levels of spatial granularity, ranging from country to city level. It is important to note that when annotating epidemiological textual data, choosing one label per article may lead to disagreements between annotators and information loss as different types of information (such as outbreak events, description of prophylactic measures, consequences, etc.) can coexist within the same document. Therefore, prioritizing information is essential to bypass this limitation.
The detailed guidelines, along with the manually annotated corpora, are valuable resources available for the community in an open source data repository [24].
In the next chapter (Chapter 6), we will evaluate EpiDCA on this dataset.

CHAPTER 6

EPIDCA EVALUATION

This chapter focuses on evaluating the proposed model, EpiDCA, across various geographical contexts and case studies.
Using the three corpora (*DB_AI_France*, *DB_ASF_Europe*, and *DB_WND_Europe*) generated through the annotation and guideline methods described in Chapter 5, we evaluate EpiDCA in terms of classification and reactivity, and sensitivity to the defined parameters.

6.1 Introduction and Objectives

EpiDCA showed promising results when applied to AI in Asia (see Chapter 4), and better results were obtained when safe signals (environmental data) were taken into account (see Chapter 4, Section 4.4). In this chapter, we want to evaluate the robustness and genericity of EpiDCA when applied to different geographical contexts and case studies.
Thus, we created an original and comprehensive dataset that includes articles related to three different epidemiological systems (AI in France, ASF, and WND in Europe) (see Chapter 5). We obtained a corpus composed of three annotated corpora including: 717 events for AI, 300 for ASF, and 409 for WND (see Chapter 5). The epidemiological metadata mentioned in the texts were extracted to generate danger signals. The locations were extracted and associated with their geographical coordinates, to project the events on risk maps and generate the safe signals.
In this chapter, we will evaluate EpiDCA in terms of classification. reactivity (meaning if events are detected timely, before the official confirmation by IBS systems), we will also evaluate the impact of spatial information granularity on the classification accuracy, and conduct a sensitivity analysis to assess how the parameters impact the method's performance.

6.2 Methodology

6.2.1 Classification methods

Parameters setting

In this section, we will present the parameters setting of danger and safe signals, followed by the temporal window and DCs coverage.

Danger signals

As defined in Chapter 4, epidemiological metadata extracted from detected articles were used to generate danger signals. Table 6.1 summarizes the parameters used to generate the danger signals for each disease.
We considered three categories of parameters for each case study: the source (official/non-official), the host, and a third parameter. For AI, the third parameter is the subtype. For WND and ASF, where no subtype is reported, we took into account the disease name. In some instances, the disease name is clearly mentioned, while in other cases it is either not specified or other diseases are mentioned.
Each parameter category is defined by a minimum and maximum score. Initially, we es-

tablished a range for each parameter category based on the literature and expert recommendations. Subsequently, experiments were conducted on a sample guided by these expert-recommended values to determine the most appropriate scores within these predefined ranges. Higher scores indicate greater relevance for epidemiological data. For example, regarding AI subtypes, HPAI subtype scores higher than LPAI and unspecified subtypes. For ASF, it has been demonstrated that in Europe, wild boars play a major role in the spread of the disease, Thus, a higher score is attributed to wild boars compared to domestic pigs [151, 152].

Defining the danger signals for WND has been more challenging due to two main factors. First, because multiple hosts (birds, humans, horses) are involved. And second, because WND is a vector-borne disease, which requires considering information on the activity of the vector (mosquitoes), and it is common that information about the vectors activity is reported on online media.
The highest scores were given to the 'human' and 'bird' labels, because surveillance and media predominantly focus on human cases of WND, and birds play a crucial role in amplifying the viremia before transmission (see Chapter 2, Section 2.4.1). Therefore, events detected in birds are considered as early signals of a potential WND outbreak.
For convenience, the label 'mosquitoes' is listed alongside hosts in the Table 6.1 however, it is important to note that mosquitoes are vectors rather than hosts.
Because we have more labels for WND compared to AI and ASF, and the overall Ds must not exceed 100 (as defined in the literature [67]), it was decided to lower the Source score to a maximum of 20. This adjustment was made because the Source parameter appeared to be the least impactful when the sensitivity analysis was conducted for the first EpiDCA evaluation (see Chapter 4, Section 4.3.5).

Parameters	**AI**	**ASF**	**WND**
Source	Official = 30 Non-Official = 20	Official = 30 Non-Official = 20	Official = 20 Non-Official = 10
Subtype /disease name	HPAI = 40 LPAI = 30 Unspecified = 10 Other = 0	ASF = 40 Unspecified = 10 Other = 0	WND = 40 Unspecified = 10 Other = 0
Host	Domestic birds = 30 Wild birds = 20 Unspecified = 10 Humans = 5 Other = 0	Wild boars = 30 Domestic pigs = 20 Unspecified = 10 Meat = 5 Other = 0	Humans = 40 Mosquitos = 30 Birds = 30 Horses = 20 Unspecified = 10 Other = 0

Table 6.1: Parameters used for AI, ASF, and WND.

Suitability maps and safe signals

Safe signals were generated following the same method described in Chapter 4, Section 4.3.2. First, we used existing recent risk maps or updated ones with recent environmental data to produce suitability maps for disease occurrence for each case study. These maps were developed using different methods, such as statistical models as described by [71], or the MCDA approach as described by [144]. Then, suitability values were converted to safe signals by applying a decreasing linear transformation. Safe signals thus lie within a range from 0 (the environment is suitable for disease occurrence) to 100 (the environment is not suitable for disease occurrence).
Next, the events were associated with their environmental data by spatial correspondence using QGIS[1], and the "point sampling tool" plugin[2] that allows one to assign to the events (points) the attributes (safe signal scores) of the underlying raster risk map.
For AI in France, we used a suitability map produced according to the methodology described in a study by [71] (see Figure 6.1). This study ranked spatial predictor variables related to poultry production based on their significance in the spatial distribution of HPAI H5N8 outbreaks during the 2016–2017 epizootic. Twelve variables, primarily concerning poultry production and water bird habitats, were evaluated for their impact on the spatial distribution of these outbreaks. We used an updated version of this map, which includes predictor variables for the year 2020-2021.

For ASF in Europe, the suitability map was created based on the ASFORCE report (Targeted Research Effort on African Swine Fever)[3] as part of an European project. Their report presents a risk map of ASF transmission from wild boars to domestic pigs in Europe, based on four risk factors; density of rural population, pig population density in the low-biosecurity sector, density or rural settlements and density of secondary roads. The layers of these four risk factors were combined using the Weighted Linear Combination (WLC) method [101]. The map produced was based on data from 2005, we updated the produced map using more recent data on domestic pigs density from 2015 available on the FAO website[4]. ASF events were then projected onto this updated map using QGIS (see Figure 6.2).

For WND in Europe, the suitability map was produced in a study conducted by [154] using the method of a statistical model. The key predictors used include: temperature anomalies in July, remotely sensed Modified Normalized Difference Water Index (MNDWI) anomalies in early June, presence of wetlands, locations of birds' migratory routes, and the occurrence of WND in the previous year. The resulting map (Figure 6.3) shows the predicted occurrence of WND in Europe and neighboring countries at a district level. The

[1] https://www.qgis.org/de/site/
[2] https://plugins.qgis.org/plugins/pointsamplingtool/
[3] https://cordis.europa.eu/project/id/311931/reporting
[4] https://www.fao.org/livestock-systems/global-distributions/pigs/en/

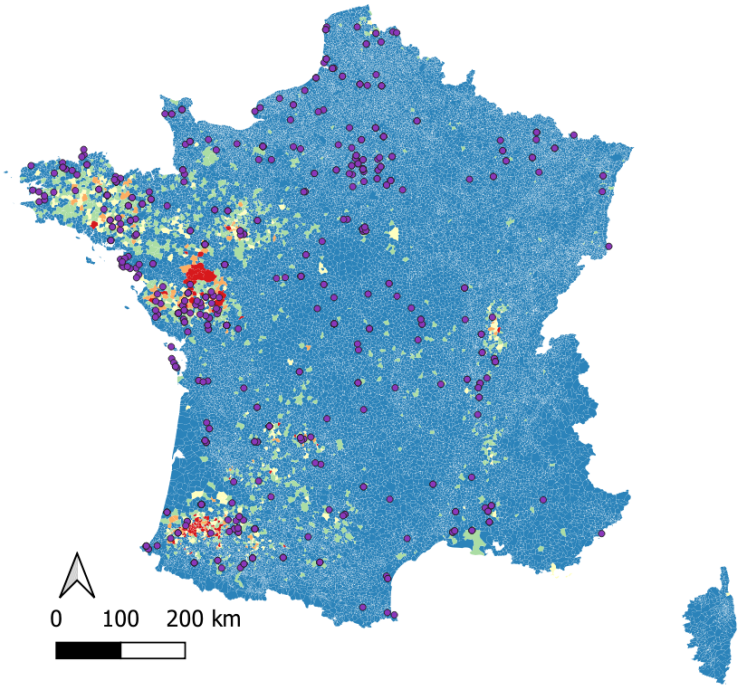


Figure 6.1: Probability of having at least one HPAI-H5N8 outbreak in sensitive hosts in France. Using 2020-2021 predictor variables. Detected AI events from *DB_AI_France* have been added using QGIS.

WND events from *DB_WND_Europe* were added using QGIS. As shown in Figure 6.2, some WND events from *DB_WND_Europe* fall outside the risk map. To avoid biases in the evaluation due to missed values within the safe signals (Ss), these events were excluded from further evaluation. As a result, the number of events that will be mentioned in the remainder of this chapter is 354 events for WND.

Temporal window and radius of coverage

These parameters were defined based on expertise and literature. For AI and ASF, the parameters were primarily based on control and surveillance measures, as discussed in Chapter 2, Sections 2.2.2 and 2.3.2. ASF is mainly transmitted by wild boars. Its natural diffusion dynamics are more easily identifiable than for AI, it occurs gradually, in a slow and continuous manner. The radius of coverage of the DCs in this case, is set to 10 km, which corresponds to the distance for which restrictions and control measures are implemented (surveillance zone) around ASF outbreaks. The migration threshold of the DCs has been sat to 40 days because, beyond this period, if no new event is detected, the affected area is considered to be free from the disease, despite the virus being resistant.
For WND the DCs coverage is set to 20 km, however the temporal window is extended to 90 days because the virus can be maintained throughout the active season of the vector [136]. A summary of the temporal and spatial parameters used for the three case studies is provided

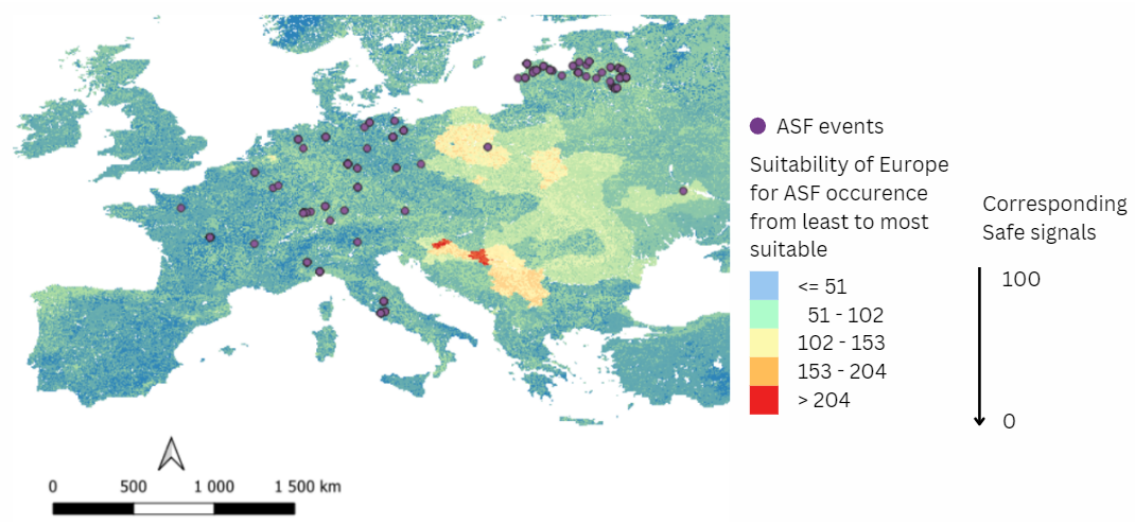


Figure 6.2: Suitability for occurrence of ASF outbreaks in domestic pigs in Europe. The map produced by [7] updated with recent data on domestic pigs. ASF events from *DB_ASF_Europe* have been added using QGIS.

in Table 6.2. To evaluate the overall classification on the three case studies, We applied the

Parameters	**AI**	**ASF**	**WND**
DCs Coverage	20 km	10 km	20 km
Migration Threshold	21 days	40 days	90 days
References	[127, 9]	[162]	[77]

Table 6.2: Spatial and temporal parameters used for AI, ASF and WND. Based on expertise and literature.

same approach as the one described in Chapter 4, Section 4.3.3. We calculated the Precision, Recall, and F-score for each of the relevant and irrelevant classes. Then, we computed the weighted F-score, which takes into account the imbalanced nature of the three corpora. This was done in two rounds: the first one, without computing the Safe signals (Safe signals set to 0), and a second round, with the Safe signals included.

6.2.2 Spatial analysis methods

Spatial analysis aims to evaluate the impact of an event's location granularity on the model's performance. Specifically, we investigate whether the model performs better when locations are provided at different levels of granularity, such as country, region, department, and city levels as described in Table 5.3, Chapter 5.

To achieve this, we filtered the dataset, as shown in Figure 6.4 and calculated the output metrics for different levels of spatial granularity. We progressively removed events asso-

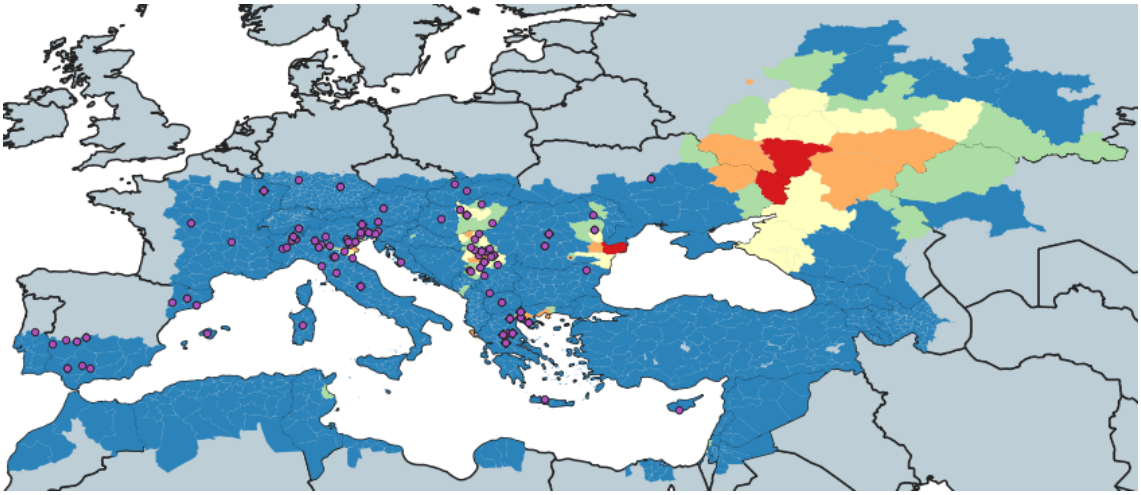


Figure 6.3: Map illustrating the predicted probability of WND occurrence in humans, Europe and neighboring countries [154]. detected events from *DB_WND_Europe* were added using QGIS.

Figure 6.4: Visualization of spatial granularity levels mentioned in the dataset. With 3 levels for *DB_AI_France* corpus and 3 levels for both *DB_ASF_Europe* and *DB_WND_Europe* corpora.

ciated with each level of location to observe the impact on the algorithm's classification performance.

6.2.3 Reactivity analysis methods

As described in Chapter 4, Section 4.3.4, the reactivity, is defined as the time difference, measured in days, between the maturation date of the DCs and the confirmation date of an event that occurred at the same location. This measure of reactivity allows us to assess whether our system is capable of early detection of outbreaks.

Due to time constraints, data availability, and the time-consuming nature of linking detected events with confirmed events from IBS systems, we were unable to include the reactivity analysis for the *_DB_AI_France* corpus in this chapter. However, the reactivity of EpiDCA to AI has already been assessed in Chapter 4 on the *DB_AI_initial* corpus.

For ASF confirmed events, we referred to EMPRES-i database. ASF events are reported

with the date of confirmation, the affected host (either wild boar or pig), and the location with its coordinates. Each row in the database corresponds to a single host, with separate entries for each case, particularly for wild boars.
For WND confirmed events, we referred to ECDC's TESSy database[5]. TESSy provides case-based data, reporting only confirmed WND cases in humans. Each row corresponds to a single case, with available information including the date of diagnosis and location details at the NUTS-3 level. NUTS stands for Nomenclature of Territorial Units for Statistics [128]. The standardization NUTS system is hierarchical, subdividing each member state into three levels: NUTS 1, NUTS 2, and NUTS 3, with each level being a further subdivision of the previous one. For example, in France, the NUTS 3 level corresponds to departments.

6.2.4 Sensitivity analysis methods

Sensitivity analyses for both ASF and WND were conducted as detailed in Chapter 4, Section 4.3.5. The first phase involved assessing the systematic evaluation of parameters, while the second phase used the Morris One-at-a-Time (OAT) method to identify the most influential parameters.

6.3 Results and discussion

In this section, we will present and discuss the results regarding the classification, spatial analysis, reactivity, and sensitivity.

6.3.1 Classification results

Classification measures were calculated on both relevant and irrelevant classes with and without including Safe signals (see Table 6.3). Overall, for these three corpora, as with *DB_AI_initial*, Epi_DCA achieved better results when considering safe signals (Ss).
This confirms that including environmental data enhances classification, with F-scores of **0.642**, **0.843**, and **0.848** for *DB_AI_France*, *DB_ASF_Europe*, and *DB_WND_Europe*, respectively.
Results are summarized in Table 6.3. EpiDCA effectively detects the relevant class despite the datasets being highly imbalanced (see Chapter 5, Table 5.3).
Results obtained for the *DB_WND_Europe* and *DB_ASF_Europe* corpora are similar to those observed with the *AI_Initial* corpus (see Chapter 1, Section 4.4). This consistency

[5] https://www.ecdc.europa.eu/en/publications-data/european-surveillance-system-tessy

across different datasets demonstrates the robustness of the method when applied to different diseases.
Regarding AI, the classification results on *DB_AI_France* with an F-score of **0.642** were lower than those obtained with *DB_AI_Initial* with an F-score of **0.823**, even though both corpora pertain to the same disease. This difference could be due to several factors:
First, the main difference between the two corpora is that *DB_AI_Initial* relies on articles detected by PADI-Web and HealthMap EBS systems, while *DB_AI_France* relies on articles detected by PADI-Web only. This makes *DB_AI_Initial* more diverse in terms of sources covered and detected events. Although *DB_AI_France* contains a larger volume of data with 717 events compared to 202 events in *DB_AI_Initial*, this increased volume could potentially introduce more noise, which might negatively affect the classification.
Secondly, AI exhibits different epidemiological patterns in Asia compared to France. For instance, Asia experiences more frequent outbreaks and continuous reporting and on a larger spatial scale, resulting in distinct and varied spatio-temporal patterns that may help the algorithm identify and distinguish between different outbreak events more effectively. In contrast, France experiences less frequent outbreaks, which can lead to less variability in spatio-temporal patterns.
In conclusion, the diversity of the *DB_AI_Initial* corpus, including a range of affected hosts and broader geographical contexts (from a continent to a country), may contribute to its better classification performance compared to the more homogeneous *AI_France* corpus.

6.3.2 Spatial analysis results

In this section, we present the spatial analysis results of the three corpora. We evaluated the impact of the events' location granularity on the model's classification.
To achieve this, we filtered the dataset as shown in Figure 6.4, and each time, calculated the output metrics for different levels of spatial granularity. For each of the relevant and irrelevant classes, we calculated Precision, Recall, and F-Score. Results are presented in Table 6.4, the last column of the table represents the weighted F-Score.

DB_AI_France

For the *DB_AI_France* corpus, the precision for the positive class is highest at the City level (**0.978**), indicating that the model is most accurate at relevant events when they are reported this granularity. However, the overall performance was best at the department level, with an F-score of **0.675**. The best precision for the negative class is achieved at the Country level (**0.212**). This high precision suggests that the model is very effective at correctly identifying irrelevant events when events are reported at this broad scale. However, this might be because many irrelevant events, which are often reported with less specific location details, are

DB_AI_France			
	Precision	Recall	F-score
Without Ss			
Relevant class	0.846	0.512	0.638
Irrelevant class	0.164	0.508	0.248
Weighted average	**0.736**	**0.511**	**0.575**
Computed Ss and Ds			
Relevant class	0.883	0.588	0.706
Irrelevant class	0.212	0.587	0.312
weighted average	**0.755**	**0.587**	**0.642**
DB_ASF_Europe			
Without Ss			
Relevant class	0.830	0.953	0.888
Irrelevant class	0.607	0.269	0.373
Weighted average	**0.802**	**0.881**	**0.833**
Computed Ss and Ds			
Relevant class	0.852	0.950	0.899
Irrelevant class	0.571	0.285	0.380
Weighted average	**0.818**	**0.88**	**0.843**
DB_WND_Europe			
Without Ss			
Relevant class	0.928	0.889	0.905
Irrelevant class	0.413	0.510	0.457
Weighted average	**0.856**	**0.839**	**0.847**
Computed Ss and Ds			
Relevant class	0.928	0.882	0.904
Irrelevant class	0.419	0.553	0.477
Weighted average	**0.862**	**0.839**	**0.848**

Table 6.3: EpiDCA classification results on *DB_AI_France*, *DB_WND_Europe*, and *DB_WND_Europe*. For each test, the evaluation metrics (precision, recall and F-score) were calculated per class. The last row indicates the macro average scores.

included at the country level. Consequently, this could artificially improve precision for the negative class but might affect the overall balance and effectiveness of the dataset. Overall, the classification performance, particularly in terms of precision, is good. Detailed results are presented in Table 6.4.

DB_ASF_Europe

For the second corpus, when events reported at the country level are included, the precision for the positive class is relatively high (**0.852**), while the precision for the negative class is moderate (**0.571**). This observation aligns with the previously noted pattern, where excluding country-level events leads to a decrease in precision for the negative class. The best

	Positive Class Results			Negative Class Results			
AI_France	**Precision**	**Recall**	**F-score**	**Precision**	**Recall**	**F-score**	**Weighted F-score**
Country	0.883	0.588	0.706	0.212	0.587	0.312	0.642
Region	0.965	0.588	0.708	0.113	0.773	0.227	0.671
Department	0.963	0.567	0.714	0.134	0.744	0.277	**0.675**
City	0.978	0.543	0.698	0.089	0.791	0.16	0.671
ASF_Europe							
Country	0.852	0.950	0.899	0.571	0.285	0.380	0.843
Region	0.753	0.986	0.854	0.666	0.076	0.137	0.839
City	0.771	0.989	0.867	0.500	0.333	0.625	**0.855**
WND_Europe							
Country	0.928	0.882	0.904	0.419	0.553	0.447	**0.848**
Region	0.967	0.569	0.717	0.116	0.75	0.201	0.68
City	0.976	0.535	0.692	0.098	0.8	0.175	0.659

Table 6.4: Classification results for the three datasets. Each row in the results table corresponds to a different level of spatial detail, starting from the country level and progressing through the region level, department level, and finally the city level. Precision, Recall, and F-score are provided for each class, followed by the weighted F-score for each level.

overall performance is observed when events are reported at the city level, with a weighted F-score of **0.855**.

DB_WND_Europe

The patterns observed for the *DB_WND_Europe* corpus differ from those seen previously. Notably, the weighted F-score drops from **0.84** to **0.6** as the corpus is filtered from the Country level to the City level. A key factor contributing to this might be the relationship between locations' granularity and events' relevance. In WND articles, the granularity of location reporting does not always correlate clearly with the relevance of the events. This is because WND articles often includes a significant amount of general or broad information that are reported at various levels of granularity. For instance, information reported at the country level may include both relevant and irrelevant events, which affects the precision and recall metrics differently compared to the corpora where location granularity is more tightly linked with the relevance of events.

Another factor that might explain the observed differences is that the WND corpus covers a relatively short period compared to the two other corpora. This shorter time-frame might lead to less representative data and could affect the overall performance metrics. Especially since spatial and temporal factors are closely linked in the method used (as discussed in Chapter 4, Section 4.3.5), the shorter reporting period might further complicate how these factors interact, potentially impacting the precision and recall metrics.

6.3.3 Reactivity analysis results

DB_ASF_Europe

The reactivity varied from 30 days before to 48 days after official confirmation. Out of the 300 DCs analyzed, 242 were classified as matured. Among these matured DCs, 17.33% (52/300) was associated with early detected events, 9% (27/300) were linked to late detected events, and 10% DCs (30/300) matured the same day as the official notification. These results confirm the effectiveness of EpiDCA in detecting timely and confirmed events,as shown in Figure 6.5.

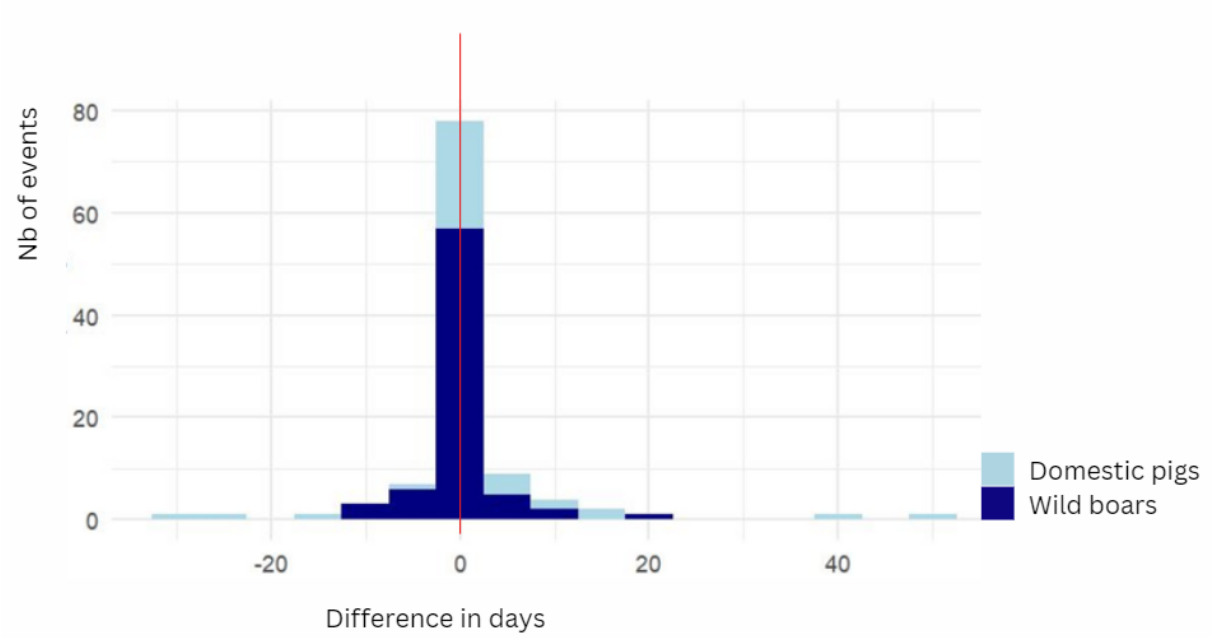


Figure 6.5: Reactivity of EpiDCA to ASF events.

Another important aspect, shown in Figure 6.5, is that early detection of events was observed in both wild boars and domestic pigs but the results does not clearly indicate whether one type of hosts is detected better than the other (unlike the clear distinction seen with AI and the early detection in wild birds). Logically, EpiDCA and EBS systems in general, might be more effective in detecting wild boars events due to passive surveillance methods that randomly uncover carcasses and affected wild boars (see Chaper 2, Section 2.3.2). However, this advantage is not distinctly apparent in the results. One reason could be the complexity of event linking: reported events in media outlets usually define outbreaks involving multiple hosts, while EMPRES-i records each ASF event per host. This inconsistency can introduce bias, as a single reported outbreak involving several hosts might be recorded as multiple separate events. Additionally, 44.33% (133/300) of the matured DCs were not associated with any confirmed event. This can be explained by several factors. It may be due to incomplete

declarations not recorded in systems like EMPRES-i, or from false alarms later refuted by laboratory results.
The primary challenge in linking events reported in media articles with the confirmed events stored in the EMPRES-i dataset lies in the differing reporting structures and levels of detail provided. Media articles typically offer general information about events without specifying exact numbers or precise locations of affected hosts. Consequently, a single article may refer to one or multiple events occurring in one or several locations.
On the other hand, the EMPRES-i dataset is case-based for ASF events, with each row representing an individual host. This is particularly relevant for wild pigs, as health authorities are required to report the exact location where carcasses are found. Consequently, it is common to have multiple rows with the same date but slightly different coordinates.
Moreover, unlike AI, where multiple indicators such as subtype, number of affected hosts, number culled, location, and date of confirmation define an event, ASF events are determined by only three factors: host, location, and confirmation date. This makes it challenging to efficiently link reported events across datasets, especially in scenarios where multiple events are reported at the same location with several days or weeks in between, as the linking process becomes even more complex.

DB_WND_Europe

We relied on the TESSy database for our analysis because the *DB_WND_Europe* corpus predominantly focused on human cases, accounting for 77% (272/354) of the events. This approach allowed us to concentrate specifically on human cases, though we did not include events involving birds, horses, or vectors that tested positive for WND.
From the analysis of human events, reactivity varied from 2 to 75 days after official confirmation. Out of the 354 DCs analyzed, 275 were classified as matured, with 52.7% (145/275) linked to a confirmed event.
This is not surprising, as the reactivity of IBS systems is typically more timely for human case confirmations due to the direct involvement of individuals in the healthcare system, such as hospitalizations or medical visits. This aligns with the results observed for human cases of AI, in Chapter 1, Section 4.4, and is supported by other studies [158, 59]. Furthermore, for WND cases in Europe, the ECDC relies on weekly updates and integrates data from the TESSy database to ensure that new human infections are promptly captured and reported.
In the future, it is crucial to assess reactivity of WND not only for human cases but also for events involving birds, horses, and mosquitoes that test positive for WND virus. Linking the locations of these events with human case locations could enhance our understanding of reactivity, as early detection in birds and other hosts is considered a significant early signal for potential outbreaks.
In addition, similar to ASF, only three factors determine a WND event (location, confirmation date, and host). This makes it difficult to link events from two different databases with

certainty, especially when multiple outbreaks occur at the same location just a few days or weeks apart.
Reactivity results showed that while linking official and non-official data is a promising approach to assess EBS reactivity, it is more challenging for certain diseases due to differing reporting configurations and the nature of the events.

6.3.4 Sensitivity analysis results

The same method described in Chapter 4 Section 4.3.5 was followed to conduct the sensitivity analysis on the corpora *WND_DB_Europe* and *ASF_DB_Europe*.
Overall, the model exhibited robustness, with minimal changes in results, and the weights assigned by the experts appeared to be the most suitable for the datasets.
When applying the OAT method, we observed that all parameters exhibited a negligible effect on the results, as indicated by their σ^* values being less than 0.1 (see Figure 6.6). This demonstrates the overall robustness of EpiDCA.
For each parameter, we made the following observations:

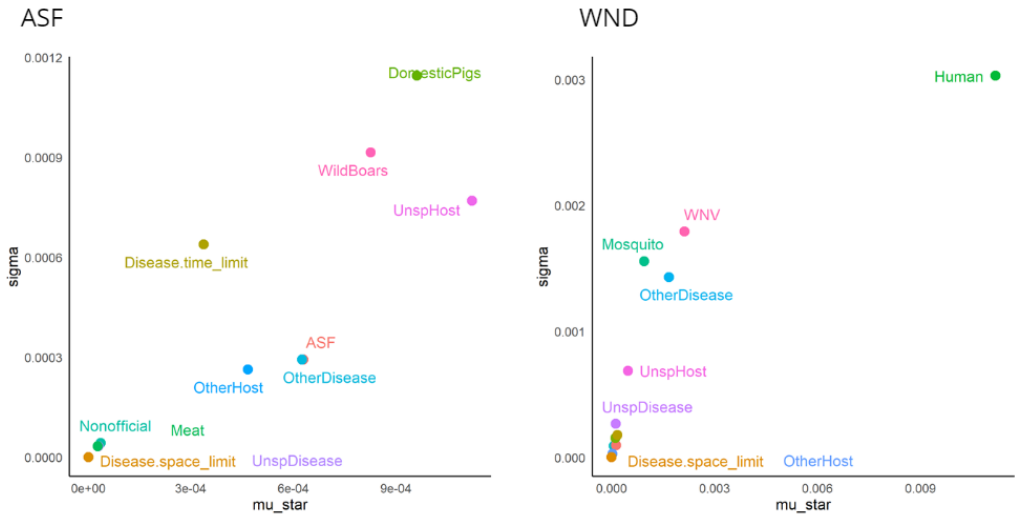


Figure 6.6: Morris OAT results for *ASF_Europe* (left), and *WND_Europe* (right). The graph represents the average of elementary effects in absolute values (μ^*) according to their standard deviation (σ) with respect to model outputs.

Same as for *AI_Initial*, results indicate that the source parameter had no impact on the classification results, this is due to the fact that the datasets used exclusively consisted of non-official sources. The "Disease name" parameter proved to be the parameter with the most significant impact for the WND corpora, and when setting this parameter to 0 the F-score dropped from 0.8 to 0.6. For the *DB_ASF_Europe* corpora, the "Disease name" parameter appeared to have a negligible impact.
For both corpora, assigning a value of 0 to the "Other disease" label proved crucial for filter-

ing out irrelevant articles that mentioned other diseases. The Host parameter group demonstrated a significant influence on the *DB_WND_Europe* corpora; specifically, when set to 0 the F-score declined from 0.8 to 0.6, Notably, the label "Human" demonstrated the highest influence, this can be explained by the prevalence of human cases in the *DB_WND_Europe* corpus. Similar pattern was observed in the *DB_ASF_Europe* corpus with the label 'Domestic pigs'.
In all three datasets, assigning a value of 0 to the "Other host" label proved crucial for filtering out irrelevant articles. This parameter played a key role in accurately classifying irrelevant events that mentioned hosts not normally affected by the studied diseases.
The temporal parameter impacted the classification results in both the AI and ASF datasets, whereas its influence was negligible in the WND dataset. This observation may be attributed to the relatively short period covered by the WND dataset.
For the spatial parameter, the same observations as described in Chapter 4, Section 4.3.5 were noted. Assessing the temporal and spatial parameters separately poses challenges due to their inherent interdependence.

6.4 Conclusion

In this chapter, we tested EpiDCA across various geographical contexts and epidemiological systems, including AI in France, WND and ASF in Europe.
Overall, the classification results demonstrated the robustness of the model across different datasets, and its ability in detecting relevant events even in highly imbalanced datasets. EpiDCA achieved very good F-scores of 0.843 and 0.848 for the *DB_ASF_Europe* and *DB_WND_Europe* corpora respectively. Although the results for the *DB_AI_France* corpus were slightly lower than those for the *DB_AI_Initial* corpus, with F-scores of 0.642 and 0.704 respectively, it is important to note that these results are still promising, and it highlight that dataset diversity, including variations in epidemiological data, spatiotemporal information, and reporting frequency, might influence the classification outcome.
The reactivity analysis revealed that linking official with non-official events is a very effective approach to assess the method's reactivity and showed very good results for early detecting AI and ASF events, and it underscores the need for a more comprehensive approach that considers both human and animal hosts to assess reactivity in multi-host and vector-borne diseases like WND. It also highlights the challenges in linking data, as this process is time-consuming and complex, especially if the datasets do not share the same structure (e.g., event-based vs. case-based datasets).
Spatial analysis showed that a minimum scale is required for optimal performance and that the way events are reported also influences classification, as there might be a correlation between the spatial granularity and the relevance of the reported events.
Moreover, the sensitivity analysis confirmed the robustness of the model's parameters.

CHAPTER 7

EXPANDING EPIDCA TO CONSIDER ADDITIONAL COVARIATES

In this final chapter, we propose an extension of the EpiDCA designed to integrate external covariates and real-time environmental data. We outline the methodology and provide an initial test using the *DB_WND_Europe* corpus. Preliminary findings are discussed, highlighting potential for future development and refinement of the approach.

7.1 Introduction and Objectives

One of the foundations of EpiDCA is the integration of environmental spatio-temporal information through safe signals derived from risk maps. These maps are created using different methods, all relying on historical environmental data, such as temperature, host population density, or case occurrences from the previous year. In the precedent chapters (Chapters 4 and 6), this approach has proven efficient, yielding better classification results when considering safe signals. However, it relies on static data and does not integrate new, real-time environmental information.

To address these limitations, we propose an extension that allows for the consideration of external, real-time covariates. This can either complement the risk maps or serve as a valuable alternative in scenarios where the risk map is unavailable.

To achieve this, we rely on the literature, particularly insights from the early versions of DCA (see Chapter 3, Section 3.4.1), which introduced the concept of "Inflammation signals" designed to amplify both Danger and Safe signals [67, 16, 32].

Our assumption is that integrating external covariates (real-time environmental data) might either enhance classification performance and/or reduce the maturation delay of the DCs (the difference between the maturation date and the creation date), thereby improving the model's reactivity.

In this chapter, we present the method for an initial test, along with preliminary results.

To conduct this initial experiment, we rely on the *DB_WND_Europe* corpus.

WND (described in Chapter 2, Section 2.4) is a vector-borne disease highly influenced by environmental drivers [136], which makes it an ideal candidate for testing how real-time environmental data impact the model's outcomes.

7.2 Methodology

The EpiDCA workflow remains unchanged and includes the following phases: Pre-processing and Categorizing, Detection, Context Assessment, and Classification (see Chapter 4, Section 4.2). However, in this extended version of EpiDCA, there is a modification in the Context Assessment phase. Specifically, we include external covariates by integrating the Inflammation signals (I), as described in the literature [32, 67, 16].

In the Context Assessment phase, the updated CSM function operates as follows:

$$CSM = ((W_D \times S_D) + (W_S \times S_S)) \times (1 + I) \tag{7.1}$$

where:

- W_D and W_S are weights for the Danger and Safe signals S_D and S_S, respectively.
- I is the inflammation signal.

The concept of the inflammation signal within the DCA is cited in the literature. However, while it has occasionally been mentioned, most proposed DCAs tend to ignore these signals and base their implementations on DSs and SSs. To the best of our knowledge, specific values and worked examples for the inflammation signal have not yet been presented.
In our proposed extension, the inflammation signal ranges between 0 and 1 and represents a normalized risk (or suitability score) derived from a given covariate. For example, if a disease has an elevated chance of transmission under certain climatic conditions, the inflammation signal would reflect this increased risk. Thus, when the signal reaches its maximum value, the global CSM is multiplied by 2.
As described in Chapter 2, Section 2.4, several research studies have investigated the environmental drivers of WND. It is noted that in Europe, key climatic drivers include the temperature, NDVI, precipitation, and migratory routes, among others. A study conducted by [54, 141] further reinforced this by demonstrating that ambient temperature plays a significant role in increasing the vectorial capacity of *Culex* mosquitoes, thereby accelerating their transmission cycle, the biting rate, and the transmission probability, which in turn results in outbreaks. Another study [141], demonstrated that transmission occurs within a range of temperatures from 12°C to 35°C, with the highest risk at an optimal temperature of 24°C. Meanwhile, [54] found that in Europe, the mean temperature of the warmest quarter ranged between 20-26°C in regions with WND occurrence.
To extend the method, an initial trial was conducted by integrating real-time temperature data.
To do so, temperature was converted into an Inflammation signal (I) that reflects the optimal temperature for WND transmission. Figure 7.1 shows that I is low at extreme temperatures and increases to a maximum value within the optimal temperature range between 20 degrees and 26 degrees.

Temperature data extraction

Daily mean temperature data were extracted from ERA5-Land, a dataset produced by the European Centre for Medium-Range Weather Forecasts (ECMWF) under the Copernicus Climate Change Service. This dataset provides high-resolution global climate data with hourly updates and a 9 km spatial resolution, covering the period from 1950 to the present, and supports comprehensive land monitoring by offering detailed insights into various environmental variables [113].
To download the temperature data, we used Google Earth Engine (GEE), a cloud-based geospatial retrieval and processing platform that allows users to explore, analyze, and down-

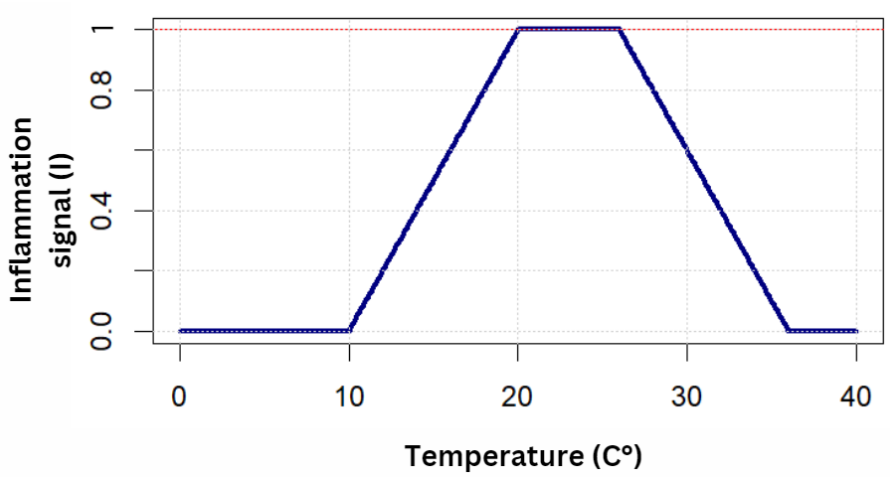


Figure 7.1: Relationship between Inflammation Signal (I) and the temperature.

load spatial data directly from a web-based editor.
To conduct a series of tests using different time intervals of temperature data, we extracted a time series of temperature data for each location in our dataset. This time series spans from the publication date, which is also the detection date of the event, to 60 days prior to the detection date. This approach ensured that we could assess the impact of temperature on our classification results across different time frames. We then linked each location with the corresponding temperature data for the dates we wanted to analyze.

Classification analysis

The classification analysis was conducted in the same way as described in Chapter 4, Section 4.3.3 and Chapter 6, Section 6.2.1. The only difference is in the Context assessment phase, where we used Eq. 7.1 to calculate the CSM values.
We conducted a series of tests considering the daily mean temperature recorded at the events' locations for various lags: d-3, d-30, d-50, and d-60. This was done to account for the delayed impact temperature can have on disease transmission.
After each test, we calculated the precision, recall, and F-score for the positive and negative classes, followed by the weighted F-score.

Reactivity analysis

In this context, we calculated the reactivity as the difference in days between the creation date and the maturation date of a DC.
We conducted this analysis with and without integrating the Inflammation signal to see if it impacts reactivity, as our hypothesis is that it might reduce the maturation delay that is

calculated for each mature DC as follow:

$$Reactivity = Maturation\_Date - Creation\_Date \tag{7.2}$$

7.3 Preliminary results and discussion

Classification results

The baseline model, which does not include temperature data, achieved a weighted F-score of **0.848**. This serves as a reference for evaluating the effect of including temperature data (see first row of Table 7.1).
When temperature data from 3 days prior to the detection dates was included, the weighted F-score dropped substantially to **0.724**. This suggests that recent temperature data might introduce noise that does not contribute positively to the classification. From an epidemiological point of view, this result is not surprising. Temperature changes over such a short period are unlikely to have a direct impact on vector activity and disease transmission, thus the observed result is expected. Using temperature data from d-30 resulted in a slight improvement over d-3, with a weighted F-score of **0.739**. The best performance among the tested lags was observed with temperature data from 50 days prior to the detection dates, achieving a weighted F-score of **0.810**. This is relatively close to the baseline performance. Including temperature data from 60 days prior to the detection date resulted in a weighted F-score of **0.710**. This decline indicates that while the 50-day lag effectively captures relevant information, extending the lag to 60 days may introduce irrelevant environmental information.
Overall, the results highlight the importance of selecting an appropriate time lag for integrating temperature data into the model. The d-50 lag appears to be the most effective, providing a balanced perspective on the environmental conditions influencing disease transmission while maintaining classification performance. The decline in performance for both shorter (d-3) and longer (d-60) lags suggests that there is an optimal window where temperature data is most relevant. However, the dataset used in this study is limited in terms of the time window (from June to September 2022), making it difficult to draw definitive conclusions, In addition, the temperature information is already partially integrated into the risk map through the variable:"Temperature anomaly for the month of June". In the case of WND, it would be particularly interesting to apply this method to a dataset covering at least a one-year period. This would allow us to evaluate whether the method can effectively filter false positives, especially given that WND cases are reported throughout the year, but the disease is characterized by seasonality and is highly influenced by temperature and other environmental factors.

	Positive Class Results			Negative Class Results			
	Precision	**Recall**	**F-score**	**Precision**	**Recall**	**F-score**	**Weighted F-score**
Baseline	0.928	0.882	0.904	0.419	0.553	0.447	**0.848**
d-3	0.92	0.679	0.782	0.228	0.617	0.333	0.724
d-30	0.922	0.702	0.797	0.241	0.617	0.347	0.739
d-50	0.923	0,823	0,87	0.325	0.553	0.409	**0.810**
d-60	0.801	0.730	0.764	0.311	0.402	0.351	0.710

Table 7.1: Classification performance metrics for different temperature lags. The first row shows the baseline results (without temperature), while the remaining rows show results obtained with various temperature lags.

Reactivity results

The reactivity was calculated as the difference in days from between the maturation date and the creation date of each DC.
Given that improving reactivity would be less meaningful if it came at the expense of classification performance, we compared results from two scenarios: the baseline that doesn't consider external covariates and the test using temperature data from 50 days prior to the detection dates (see Table 7.1).
The reactivity analysis results (see Table 7.2) showed that including temperature in the analysis reduced the maximum delay from **30 days** to **26 days** and the mean delay from **5 days** to **4 days**. This confirms that considering real-time data can be a good approach to enhance the model's reactivity.

	Weighted F-score	**Mean reactivity delay (Days)**	**Delay (max nb of days)**
Baseline	**0.848**	5	30
d-50	0.810	**4**	**26**

Table 7.2: Comparison of the reactivity resluts with and without including the temperature.

7.4 Conclusion

In this chapter, we presented an extension of EpiDCA designed to integrate external real-time environmental data into the model. We achieved this by integrating inflammation signals into the Context assessment phase, as described in the literature. An initial test was conducted using the *DB_WND_Europe* corpus and daily mean temperature data for various time lags relative to the detected events' locations and dates.
Our results show that integrating temperature data affects both classification performance and reactivity, and underscore the importance of relying on expert knowledge to select the appropriate variable and time windows for integrating real-time environmental data.

In terms of reactivity, including temperature data improved the model's reactivity by reducing both the maximum and mean delays between the creation and maturation of DCs. Specifically. This demonstrates that integrating temperature data can improve the model's reactivity without significantly compromising classification performance.
However, while our extension of EpiDCA shows promising results, further tests with more extensive datasets and the inclusion of additional covariates are necessary to fully evaluate benefits and limitations of integrating real-time environmental data in the model.
Additionally, we can consider developing a more sophisticated CSM function than the one used here, which, as described in the literature, defines the I signal as a value between 0 and 1, implying it is always additive. It might be interesting to explore a calculation method that, in certain instances, could reduce the global CSM, potentially offering a more nuanced and effective approach.
The extension of EpiDCA offers various perspectives for methodological improvement. First, this approach needs to be tested on a more extensive dataset and include additional covariates to assess its impact over a larger time window. Second, it would be interesting to explore a more sophisticated approach to both the function that converts covariates to the I signal and the global CSM function.

CHAPTER 8

CONCLUSION AND PERSPECTIVES

In this final chapter, we present our general conclusions and perspectives. We begin by summarizing the key contributions. Following this, we explore potential future research directions and applications of our work.

8.1 Summary of the main contributions

This work falls within the context of event-based surveillance (EBS) and sits at the intersection of epidemiology and computer science.
The main objective of this thesis was to develop a comprehensive approach for EBS systems that goes beyond traditional text-based classification methods by considering the environmental context of the detected events.
In this thesis the four main contributions are:

1. Development of an unsupervised method called EpiDCA
2. Creation of an annotation method and production of an original dataset.
3. Evaluation of EpiDCA on the produced dataset.
4. Introduction of an extension of EpiDCA

Developement of EpiDCA

EpiDCA, which is an adaptation of the Dendritic Cells Algorithm (DCA) [67] inspired by danger theory [105], introduces a new approach to EBS systems. To the best of our knowledge, this is the first application of this kind in the context of EBS surveillance. The methodological contribution of EpiDCA lies in its innovative approach to overcome specific limitations of the original DCA, including the integration of spatio-temporal information during the detection phase. We successfully applied EpiDCA to a first case study using the PADI-Web and HealthMap dataset, which includes documents related to avian influenza (AI) in Asia from 2018 to 2019. We used environmental data derived from an updated suitability map for AI in the same region. The model's performance was evaluated using precision, recall, and F-score metrics, achieving an F-score of 0.823 on an imbalanced dataset and 0.90 on a balanced dataset. EpiDCA was also compared with leading supervised machine learning methods and demonstrated competitive performance. These results highlight the effectiveness of considering disease risk factors in event classification.
Additionally, reactivity analysis showed that EpiDCA effectively detected outbreak events timely, particularly for wild birds.

Creation of an annotation method and production of an original dataset

As a significant contribution, we developed a comprehensive annotation method and guidelines, that was used to produce an original dataset. In addition to the annotation of the

documents by relevance, we extracted the events' epidemiological data and locations information. This dataset was specifically designed for evaluating EpiDCA and can also be used for fine-tuning language models, and training supervised machine-learning methods. Additionally, it can be used for testing spatial-based methods, as it provides epidemiological events at different granularities, from country to city levels. Importantly, this dataset, along with the annotation guidelines, is available to the community, offering a valuable resource for future studies and model development.

Evaluation of EpiDCA on the produced dataset

To ensure that EpiDCA is a robust and generic method, we evaluated it across a range of datasets representing different diseases and geographical contexts. Specifically, we tested EpiDCA on datasets for avian influenza (AI) in France, West Nile virus Disease (WND), and African swine fever (ASF) in Europe.
Our evaluation covered several aspects: classification accuracy, reactivity, spatial analysis, and sensitivity analysis. The classification and sensitivity analysis results demonstrated the robustness of the model, showing strong performance across various datasets. Importantly, considering environmental data (through the use of Safe Signals) consistently enhanced results across all configurations. Reactivity analysis demonstrated the model's capability to detect outbreaks early. Despite this analysis being more challenging in some cases, such as with WND and ASF.

Extension of EpiDCA

We proposed an extension of EpiDCA designed to integrate real-time environmental data into the model aiming to improve the classification and/or the reactivity. Drawing on literature and the concept of inflammation signals, we included temperature data into the CSM function in the Detection phase. We conducted an initial test using the WND case study, evaluating the impact of temperature data on both classification performance and reactivity. Our results indicated that while recent temperature data can introduced noise, using data from a 50-day window preserved classification performance and improved reactivity by reducing delays in DC maturation. These findings highlight the potential for enhancing EpiDCA with real-time environmental data, though further testing with larger datasets and additional variables is needed to fully assess its effectiveness and limitations.

8.2 Perspectives

This thesis is interdisciplinary, combining data-driven approaches with model-based methods that integrate expert knowledge. This original work opens new perspectives for various applications and methodological contributions in the field of epidemiological surveillance, which we discuss in this section.

EpiDCA is an adaptation of the DCA. While it uses the same phases: Pre-processing and Categorization, Detection phase, Context Assessment phase, and Classification, this version addresses limitations related the Detection and Context Assessment phases highlighted by previous studies [68, 32, 55].

One noted limitation in EpiDCA is that the Pre-processing and Categorization phase involved the manual extraction of epidemiological data (Danger signals) and locations from the text, which was time-consuming and could be improved. Several studies have addressed these limitations, ranging from DCA studies proposing optimization methods [49, 33] to text-mining works that focus on the automated information extraction from texts [155, 148]. Combining these advances could enhance EpiDCA's efficiency and usability.

In addition to these improvements, another significant contribution of EpiDCA has been the consideration of the spatio-temporal information of the detected events in the Detection phase. However, integrating network analysis methods into EpiDCA could offer additional layers of understanding. For example, by identifying highly connected holdings that are critical for surveillance and disease prevention [45], rather than focusing on the geographical distance only.

Integrating environmental data through risk mapping appears to be a promising approach. Various tools can provide this input, and it is important to combine available data sources. For example, Arbocarto [104] is an operational spatial modeling tool that predicts the dynamics of *Aedes* mosquito species based on weather and environmental variables. It provides weekly updates on the distribution of these mosquitoes, which are vectors for topical diseases such as dengue fever, Zika virus, and Chikungunya.

Moreover, to convert the risk map information (such as risk index or probability of occurrence of a given disease) into a Safe signal, we used a relatively simple linear function. However, it may be beneficial to explore alternative methods that could provide more nuanced insights. For example, applying an exponential decay function, which has been effectively used in epidemiological modeling [21].

In this work, we have successfully applied EpiDCA to three distinct case studies, demonstrating its robustness and genericity in animal and human disease surveillance. Looking ahead, we can extend this approach by adapting and applying it to other areas that rely on similar data sources (textual and environmental data). For instance, applying EpiDCA in the context of food security surveillance and plant disease surveillance, both of which are relevant and current topics in early surveillance systems [40, 92, 132].

BIBLIOGRAPHY

[1] Uwe Aickelin, Dipankar Dasgupta, and Feng Gu. "Artificial immune systems". *Search Methodologies: Introductory Tutorials in Optimization and Decision Support Techniques*. Springer, 2013, pp. 187–211.

[2] Andrea Ammon and D Faensen. "Surveillance of infectious diseases at the EU level". *Bundesgesundheitsblatt-Gesundheitsforschung-Gesundheitsschutz* 52 (2009), pp. 176–182.

[3] Marisa Arias et al. "African swine fever". *Trends in emerging viral infections of swine* (2002), pp. 119–124.

[4] Elena Arsevska. "Élaboration d'une méthode semi-automatique pour l'identification et le traitement des signaux d'émergence pour la veille internationale sur les maladies animales infectieuses". PhD thesis. Université Paris Saclay (COmUE), 2017.

[5] Elena Arsevska et al. "Monitoring disease outbreak events on the web using text-mining approach and domain expert knowledge". *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 2016.

[6] Elena Arsevska et al. "Web monitoring of emerging animal infectious diseases integrated in the French Animal Health Epidemic Intelligence System". *PLoS One* 13.8 (2018).

[7] ASFORCE Project. *ASFORCE Report (Targeted Research Effort on African Swine Fever)*. https://cordis.europa.eu/project/id/311931/reporting. Accessed: 2024-08-04. 2014.

[8] Augsburger. *Tote Rehe: Neuer Zaun im Nationalpark ist fertig*. 2022. URL: https://english.news.cn/20220829/444b6757fa714034a9e6ea33a092bbc0/c.html.

[9] European Food Safety Authority et al. "Avian influenza overview December 2020–February 2021". *Efsa Journal* 19.3 (2021).

[10] European Food Safety Authority et al. "Avian influenza overview december 2022–march 2023". *EFSA Journal* 21.3 (2023).

[11] S Arunmozhi Balajee et al. "The practice of event-based surveillance: concept and methods". *Global Security: Health, Science and Policy* 6.1 (2021), pp. 1–9.

[12] Philippe Barboza. "Evaluation des systèmes d'intelligence épidémiologique appliqués à la détection précoce des maladies infectieuses au niveau mondial." PhD thesis. Université Pierre et Marie Curie-Paris VI, 2014.

[13] Philippe Barboza 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 (2014).

[14] Christopher M Barker. "Models and surveillance systems to detect and predict West Nile virus outbreaks". *Journal of Medical Entomology* 56.6 (2019).

[15] Kazim Beebeejaun et al. “Evaluation of national event-based surveillance, Nigeria, 2016–2018”. *Emerging Infectious Diseases* 27.3 (2021), p. 694.

[16] Mourad Belhadj. “Sécurité des réseaux informatiques basée sur la théorie de danger”. PhD thesis. Université Biskra, 2012.

[17] Hannes Bergmann et al. “A review of environmental risk factors for African Swine Fever in European wild boar”. *Animals* 11.9 (2021).

[18] Nicky Best, Sylvia Richardson, and Andrew Thomson. “A comparison of Bayesian spatial models for disease mapping”. *Statistical methods in medical research* 14.1 (2005), pp. 35–59.

[19] Bradley J Blitvich. “Transmission dynamics and changing epidemiology of West Nile virus”. *Animal Health Research Reviews* 9.1 (2008), pp. 71–86.

[20] D Katterine Bonilla-Aldana et al. “Coronavirus infections reported by ProMED, february 2000–january 2020”. *Travel Medicine and Infectious Disease* 35 (2020).

[21] Milen Borisov and Svetoslav Markov. “The two-step exponential decay reaction network: analysis of the solutions and relation to epidemiological SIR models with logistic and Gompertz type infection contact patterns”. *Journal of Mathematical Chemistry* 59 (2021), pp. 128–131.

[22] Bahdja Boudoua. “An annotated Avian Influenza dataset from two event-based surveillance systems”. Version V1. *Recherche Data Gouv* (2023). DOI: 10.57745/6R81RT. URL: https://doi.org/10.57745/6R81RT.

[23] Bahdja Boudoua and Annelise Tran. “Suitability map for Avian influenza, Asia”. Version V1 (2023). DOI: 10.18167/DVN1/FYWDOJ. URL: https://doi.org/10.18167/DVN1/FYWDOJ.

[24] Bahdja Boudoua et al. “Annotated datasets from PADI-web for event-based surveillance of Avian Influenza, African Swine Fever, and West-Nile Virus Disease”. Version V1. *Recherche Data Gouv* (2023). DOI: 10.57745/99SNOZ. URL: https://doi.org/10.57745/99SNOZ.

[25] John S Brownstein et al. “Surveillance Sans Frontieres: Internet-based emerging infectious disease intelligence and the HealthMap project”. *PLoS medicine* 5.7 (2008), p. 151.

[26] GK Bruckner. “The role of the World Organisation for Animal Health (OIE) to facilitate the international trade in animals and animal products: policy and trade issues”. *Onderstepoort Journal of Veterinary Research* 76.1 (2009), pp. 141–146.

[27] Macfarlane Burnet. “Auto-immune disease: I. Modern immunological concepts”. *British medical journal* 2.5 (1959), p. 645.

[28] Ilaria Capua and Dennis J Alexander. “Avian influenza infections in birds–a moving target”. *Influenza and other respiratory viruses* 1.1 (2007), pp. 11–18.

[29] Malwina Carrion and Lawrence C Madoff. “ProMED-mail: 22 years of digital surveillance of emerging infectious diseases”. *International health* 9.3 (2017), pp. 177–183.

[30] Nigel Chaffey. *Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. and Walter, P. Molecular biology of the cell. 4th edn.* 2003.

[31] Caren Chancey et al. "The global ecology and epidemiology of West Nile virus". *BioMed research international* 2015.1 (2015).

[32] Zeineb Chelly and Zied Elouedi. "A survey of the dendritic cell algorithm". *Knowledge and Information Systems* 48.3 (2016), pp. 505–535.

[33] Zaineb Chelly Dagdia and Zied Elouedi. "A hybrid fuzzy maintained classification method based on dendritic cells". *Journal of Classification* 37.1 (2020), pp. 18–41.

[34] Erika Chenais et al. "Epidemiological considerations on African swine fever in Europe 2014–2018". *Porcine health management* 5.1 (2019), p. 6.

[35] Nigel Collier et al. "BioCaster: detecting public health rumors with a Web-based text mining system". *Bioinformatics* 2.24 (2008).

[36] Andrew A Cunningham, Peter Daszak, and James LN Wood. "One Health, emerging infectious diseases and wildlife: two decades of progress?" *Philosophical Transactions of the Royal Society B: Biological Sciences* 372.1725 (2017).

[37] Gwenaëlle Dauphin. "WHO/FAO/OIE tripartite coordination for the control and prevention of zoonotic influenza viruses. Example of OFFLU, global network of veterinary expertise". *Bulletin de l'Académie Vétérinaire de France* 168.3 (2015), pp. 224–232.

[38] Angel N Desai et al. "Infectious disease outbreaks among forcibly displaced persons: an analysis of ProMED reports 1996–2016". *Conflict and health* 14 (2020), pp. 1–10.

[39] K Devaraja et al. "Epidemiological Triad of COVID-Associated Mucormycosis and the ABCD of its Management". *Indian journal of public health* 66.4 (2022), pp. 520–521.

[40] Emily Diemer. "Leveraging Event-Based Surveillance Data to Inform Food Safety Emergency Preparedness and Response". PhD thesis. Tufts University, Gerald J. and Dorothy R. Friedman School of Nutrition ..., 2023.

[41] Marie Dion, Philip AbdelMalik, and Abla Mawudeku. "Big data: big data and the global public health intelligence network (GPHIN)". *Canada communicable disease report* 41.9 (2015), p. 209.

[42] Centers for Disease Control and Prevention. *Influenza*. https://wwwnc.cdc.gov/travel/yellowbook/2024/infections-diseases/influenza. In *CDC Yellow Book 2024: Health information for international travel*. 2024.

[43] Son Doan, Mike Conway, and Nigel Collier. "An empirical study of sections in classifying disease outbreak reports". *Web-Based Applications in Healthcare and Biomedicine* (2010), pp. 47–58.

[44] Chenlin Duan et al. "An overview of avian influenza surveillance strategies and modes". *Science in One Health* (2023).

[45] Caroline Dubé et al. “Introduction to network analysis and its implications for animal disease modelling”. *Revue Scientifique et Technique-OIE* 30.2 (2011), p. 425.

[46] ECDC. *African Development Bank turns to hedge fund to offset risk*. 2022. URL: https://www.ecdc.europa.eu/en/publications-data/west-nile-virus-europe-2022-human-cases-updated-22-september-2022.

[47] European Food Safety Authority (EFSA) et al. “Epidemiological analysis of African swine fever in the European Union during 2022”. *EFSA Journal* 21.5 (2023).

[48] elDiarioes. *Una investigación demuestra la circulación del Virus del Nilo en pájaros de Badajoz*. 2022. URL: https://english.news.cn/20220829/444b6757fa714034a9e6ea33a092bbc0/c.html.

[49] Noe Elisa, Longzhi Yang, and Nitin Naik. “Dendritic cell algorithm with optimised parameters using genetic algorithm”. *2018 IEEE Congress on Evolutionary Computation (CEC)*. IEEE. 2018, pp. 1–8.

[50] Marcos A Espinal and Prabhjot Singh. “Methods in Surveillance and Monitoring and Evaluation”. *Global Health Essentials*. Springer, 2023, pp. 549–555.

[51] European Centre for Disease Prevention and Control. *West Nile Fever - Surveillance and Disease Data*. https://www.ecdc.europa.eu/en/west-nile-fever/surveillance-and-disease-data/disease-data-ecdc. Accessed: 2024-07-29. 2024.

[52] Sana Eybpoosh et al. “Epidemiology of West Nile Virus in the Eastern Mediterranean region: A systematic review”. *PLoS neglected tropical diseases* 13.1 (2019).

[53] Li-Qun Fang et al. “Environmental factors contributing to the spread of H5N1 avian influenza in mainland China”. *PloS one* 3.5 (2008).

[54] Zia Farooq et al. “Artificial intelligence to predict West Nile virus outbreaks with eco-climatic drivers”. *The Lancet Regional Health–Europe* 17 (2022).

[55] Ehsan Farzadnia, Hossein Shirazi, and Alireza Nowroozi. “A new intrusion detection system using the improved dendritic cell algorithm”. *The Computer Journal* 64.8 (2021).

[56] Céline Faverjon et al. “Risk-based early detection system of African Swine Fever using mortality thresholds”. *Transboundary and Emerging Diseases* 68.3 (2021).

[57] Clark C Freifeld et al. “HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports”. *Journal of the American Medical Informatics Association* 15.2 (2008), pp. 150–157.

[58] C Gallardo, J Fernández-Pinero, and MJVR Arias. “African swine fever (ASF) diagnosis, an essential tool in the epidemiological investigation”. *Virus research* 271 (2019).

[59] Iris Ganser. *Evaluation of event-based internet biosurveillance for multi-regional detection of seasonal influenza onset*. McGill University (Canada), 2020.

[60] Mullusew Gashaw. "A review on avian influenza and its economic and public health impact". *Int J Vet Sci Technol* 4.1 (2020), pp. 15–27.

[61] Natasha N Gaudreault et al. "African swine fever virus: an emerging DNA arbovirus". *Frontiers in veterinary science* 7 (2020), p. 215.

[62] Gian Franco Gensini, Andrea Alberto Conti, and Donatella Lippi. "The contributions of Paul Ehrlich to infectious disease". *Journal of Infection* 54.3 (2007), pp. 221–224.

[63] Vincenzo Gervasi et al. "Evaluation of the efficiency of active and passive surveillance in the detection of African swine fever in wild boar". *Veterinary sciences* 7.1 (2019), p. 5.

[64] Jhony H Giraldo and Thierry Bouwmans. "On the minimization of Sobolev norms of time-varying graph signals: Estimation of new Coronavirus disease 2019 cases". *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE. 2020, pp. 1–6.

[65] Rohan Goel et al. "EpidNews: Extracting, exploring and annotating news for monitoring animal diseases". *Journal of Computer Languages* 56 (2020).

[66] Moisés Gonzálvez et al. "Monitoring the epidemic of West Nile virus in equids in Spain, 2020–2021". *Preventive Veterinary Medicine* 217 (2023).

[67] Julie Greensmith. "The dendritic cell algorithm". PhD thesis. Citeseer, 2007.

[68] Julie Greensmith and Uwe Aickelin. "The deterministic dendritic cell algorithm". *International conference on artificial immune systems*. Springer. 2008, pp. 291–302.

[69] Julie Greensmith, Jamie Twycross, and Uwe Aickelin. "Dendritic cells for anomaly detection". *2006 IEEE international conference on evolutionary computation*. IEEE. 2006, pp. 664–671.

[70] Feng Gu, Julie Greensmith, and Uwe Aickelin. "Theoretical formulation and analysis of the deterministic dendritic cell algorithm". *Biosystems* 111.2 (2013), pp. 127–135.

[71] Claire Guinat et al. "Duck production systems and highly pathogenic avian influenza H5N8 in France, 2016–2017". *Scientific Reports* 9.1 (2019).

[72] Kishor Datta Gupta and Dipankar Dasgupta. "Negative selection algorithm research and applications in the last decade: A review". *IEEE Transactions on Artificial Intelligence* 3.2 (2021), pp. 110–128.

[73] Ramona Alikiiteaga Gutiérrez et al. "A (H5N1) virus evolution in South East Asia". *Viruses* 1.3 (2009), pp. 335–361.

[74] Berna Haktanirlar Ulutas and Sadan Kulturel-Konak. "A review of clonal selection algorithm and its applications". *Artificial Intelligence Review* 36 (2011), pp. 117–138.

[75] Alan J Hay and John W McCauley. "The WHO global influenza surveillance and response system (GISRS)—a future perspective". *Influenza and other respiratory viruses* 12.5 (2018), pp. 551–557.

[76] M Hazelton. "Kernel smoothing methods". *Handbook of spatial epidemiology. Boca Raton: Chapman & Hall/CRC* (2016), pp. 195–207.

[77] A Hess, JK Davis, and MC Wimberly. "Identifying environmental risk factors and mapping the distribution of West Nile virus in an endemic region of North America". *GeoHealth* 2.12 (2018), pp. 395–409.

[78] Steven J Hoffman and Sarah L Silverberg. "Delays in global disease outbreak responses: lessons from H1N1, Ebola, and Zika". *American journal of public health* 108.3 (2018), pp. 329–333.

[79] Andrew G Huff et al. "Evaluation and verification of the global rapid identification of threats system for infectious diseases in textual data sources". *Interdisciplinary perspectives on infectious diseases* 2016.1 (2016).

[80] Nahla Khamis Ibrahim. "Epidemiologic surveillance for controlling Covid-19 pandemic: types, challenges and implications". *Journal of infection and public health* 13.11 (2020).

[81] Carmen Iscaro et al. "Analysis of surveillance and prevention plan for African Swine Fever in Italy in 2020". *Veterinary medicine and science* 8.4 (2022).

[82] Md Zahidul Islam et al. "A semantics aware random forest for text classification". *Proceedings of the 28th ACM international conference on information and knowledge management*. 2019.

[83] Charles A Janeway Jr et al. "Principles of innate and adaptive immunity". *Immunobiology: The Immune System in Health and Disease. 5th edition*. Garland Science, 2001.

[84] T Jacob John and Rajeev Zachariah Kompithra. "Eco-epidemiology triad to explain infectious diseases". *Indian Journal of Medical Research* 158.2 (2023), pp. 107–112.

[85] Rimjhim Kanaujia et al. "Avian influenza revisited: Concerns and constraints". *Virus-Disease* 33.4 (2022), pp. 456–465.

[86] Juhyeon Kim and Insung Ahn. "Infectious disease outbreak prediction using media articles with machine learning models". *Scientific reports* 11.1 (2021).

[87] Yun Ji Kim, Weonwoo Nam, and Jongsoo Lee. "Multiclass anomaly detection for unsupervised and semi-supervised data based on a combination of negative selection and clonal selection algorithms". *Applied Soft Computing* 122 (2022).

[88] Nicholas E. Kman and Daniel J. Bachmann. "Biosurveillance: A Review and Update". *Advances in Preventive Medicine* 2012 (2012). DOI: 10.1155/2012/301408. URL: https://doi.org/10.1155/2012/301408.

[89] Anna Kuehne et al. "Event-based surveillance at health facility and community level in low-income and middle-income countries: a systematic review". *BMJ global health* 4.6 (2019).

[90] Eric F Lambin et al. "Pathogenic landscapes: interactions between land, people, disease vectors, and their animal hosts". *International journal of health geographics* 9.1 (2010), pp. 1–13.

[91] Minsuk Lee, Weiqing Wang, and Hong Yu. "Exploring supervised and unsupervised methods to detect topics in biomedical text". *BMC bioinformatics* 7 (2006), pp. 1–11.

[92] Sue Han Lee et al. "New perspectives on plant disease characterization based on deep learning". *Computers and Electronics in Agriculture* 170 (2020).

[93] Gaël Lejeune et al. "Multilingual event extraction for epidemic detection". *Artificial intelligence in medicine* 65.2 (2015), pp. 131–143.

[94] Sunghoon Lim, Conrad S Tucker, and Soundar Kumara. "An unsupervised machine learning model for discovering latent infectious diseases using social media data". *Journal of biomedical informatics* 66 (2017), pp. 82–94.

[95] David Limon-Cantu and Vicente Alarcon-Aquino. "Multiresolution dendritic cell algorithm for network anomaly detection". *PeerJ Computer Science* 7 (2021).

[96] Jens P Linge et al. "Media monitoring of public health threats with medisys". *C. WILLIAM, CWR. WEB-STER, D. BALAHUR, et al* (2012), pp. 17–31.

[97] Valérie R Louis et al. "Modeling tools for dengue risk mapping-a systematic review". *International journal of health geographics* 13.1 (2014), pp. 1–15.

[98] Kai Ma et al. "Extraction of temporal information from social media messages using the BERT model". *Earth Science Informatics* 15.1 (2022), pp. 573–584.

[99] Lawrence C Madoff and Annie Li. "Web-based surveillance systems for human, animal, and plant diseases". *Microbiology Spectrum* 2.1 (2014), pp. 10–11.

[100] Avi Magid, Anat Gesser-Edelsburg, and Manfred S Green. "The role of informal digital surveillance systems before, during and after infectious disease outbreaks: a critical analysis". *Defence Against Bioterrorism: Methods for Prevention and Control*. Springer. 2018, pp. 189–201.

[101] Jacek Malczewski. "On the use of weighted linear combination method in GIS: common and best practice approaches". *Transactions in GIS* 4.1 (2000), pp. 5–22.

[102] Matteo Marcantonio et al. "Identifying the environmental conditions favouring West Nile virus outbreaks in Europe". *PloS one* 10.3 (2015).

[103] Monica Marchino et al. "Process evaluation of integrated West Nile virus surveillance in northern Italy: an example of a One Health approach in public health policy". *Evaluation and Program Planning* 89 (2021).

[104] Renaud Marti et al. "ARBOCARTO: an operational spatial modeling tool to predict the dynamics of Aedes mosquito species from weather and environmental variables". ESA. 2022.

[105] Polly Matzinger. "The danger model: a renewed sense of self". *science* 296.5566 (2002), pp. 301–305.

[106] Abla Mawudeku and Michael Blench. "Global public health intelligence network (GPHIN)". *Proceedings of Machine Translation Summit X: Invited papers*. 2005.

[107] Alex F McCalla. “FAO in the Changing Global Landscape”. *Department of Agricultural and Resource Economics University of California, Davis* (2007).

[108] Emily McDonald. “Surveillance for West Nile virus disease—United States, 2009–2018”. *MMWR. Surveillance Summaries* 70 (2021).

[109] Mary L McHugh. “Interrater reliability: the kappa statistic”. *Biochemia medica* 22.3 (2012), pp. 276–282.

[110] Edmond Menya et al. “Explainable epidemiological thematic features for event based disease surveillance”. *Expert Systems with Applications* 250 (2024).

[111] Mohamad Farhan Mohamad Mohsin, Azuraliza Abu Bakar, and Abdul Razak Hamdan. “Outbreak detection model based on danger theory”. *Applied soft computing* 24 (2014), pp. 612–622.

[112] Ahmed Mostafa et al. “Zoonotic potential of influenza A viruses: a comprehensive overview”. *Viruses* 10.9 (2018), p. 497.

[113] Joaquín Muñoz-Sabater et al. “ERA5-Land: A state-of-the-art global reanalysis dataset for land applications”. *Earth system science data* 13.9 (2021).

[114] Stephen Mutuvi et al. “Multilingual epidemiological text classification: a comparative study”. *COLING, International Conference on Computational Linguistics*. 2020.

[115] Eric Mykhalovskiy and Lorna Weir. “The Global Public Health Intelligence Network and early warning outbreak detection: a Canadian contribution to global public health”. *Canadian journal of public health* 97 (2006), pp. 42–44.

[116] Noele P Nelson et al. “Event-based internet biosurveillance: relation to epidemiological observation”. *Emerging themes in epidemiology* 9 (2012), pp. 1–13.

[117] NP Nelson, JS Brownstein, and DM Hartley. “Event-based biosurveillance of respiratory disease in Mexico, 2007–2009: connection to the 2009 influenza A (H1N1) pandemic?” *Eurosurveillance* 15.30 (2010).

[118] Minh-Tien Nguyen et al. “Understanding Transformers for Information Extraction with Limited Data”. *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*. 2020, pp. 478–487.

[119] Imbi Nurmoja et al. “Epidemiological analysis of the 2015–2017 African swine fever outbreaks in Estonia”. *Preventive veterinary medicine* 181 (2020).

[120] Robert Oates, Graham Kendall, and Jonathan M Garibaldi. “Frequency analysis for dendritic cell population tuning”. *Evolutionary Intelligence* 1 (2008), pp. 145–157.

[121] World Health Organization et al. “A guide to establishing event-based surveillance” (2008).

[122] World Health Organization et al. *Early detection, assessment and response to acute public health events: implementation of early warning and response with a focus on event-based surveillance: interim version*. Tech. rep. World Health Organization, 2014.

[123] Christoph Paquet et al. “Epidemic intelligence: a new framework for strengthening disease surveillance in Europe”. *Eurosurveillance* 11.12 (2006), pp. 5–6.

[124] Mathilde C Paul et al. “Quantitative assessment of a spatial multicriteria model for highly pathogenic avian influenza H5N1 in Thailand, and application in Cambodia”. *Scientific Reports* 6.1 (2016).

[125] Shlomit Paz et al. “Permissive summer temperatures of the 2010 European West Nile fever upsurge”. *PloS one* 8.2 (2013).

[126] Kim M Pepin et al. “Ecological drivers of African swine fever virus persistence in wild boar populations: Insight for control”. *Ecology and Evolution* 10.6 (2020).

[127] Maria Pittman and A Laddomada. “Legislation for the control of avian influenza in the European Union”. *Zoonoses and public health* 55.1 (2008), pp. 29–36.

[128] Meri Raggi et al. “A classification of european NUTS 3 regions”. *Publications Office of the European Union* (2013).

[129] Nasir Rashid et al. “Artificial immune system–negative selection classification algorithm (NSCA) for four class electroencephalogram (EEG) signals”. *Frontiers in human neuroscience* 12 (2018), p. 439.

[130] Colin R Reeves. “Genetic algorithms”. *Handbook of metaheuristics* (2010), pp. 109–139.

[131] Maud Reveilhac and Davide Morselli. “Dictionary-based and machine learning classification approaches: a comparison for tonality and frame detection on Twitter data”. *Political Research Exchange* 4.1 (2022).

[132] Mathieu Roche et al. “PADI-web for Plant Health Surveillance”. *International Conference on Advanced Information Systems Engineering*. Springer. 2024, pp. 148–156.

[133] Agnès Rortais et al. “MedISys: An early-warning system for the detection of re-emerging food-and feed-borne hazards”. *Food Research International* 43.5 (2010).

[134] Sihem Sahnoun and Gaël Lejeune. “Multilingual epidemic event extraction: From simple classification methods to open information extraction (OIE) and ontology”. *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*. 2021.

[135] Henrik Salje, Derek AT Cummings, and Justin Lessler. “Estimating infectious disease transmission distances using the overall distribution of cases”. *Epidemics* 17 (2016), pp. 10–18.

[136] Priya Sampathkumar. “West Nile virus: epidemiology, clinical presentation, diagnosis, and prevention”. *Mayo Clinic Proceedings*. Vol. 78. 9. Elsevier. 2003.

[137] Erika R Schwarz and Maureen T Long. “Comparison of West Nile virus disease in humans and horses: exploiting similarities for enhancing syndromic surveillance”. *Viruses* 15.6 (2023).

[138] Aakanksha Sharaff et al. “Spam message detection using Danger theory and Krill herd optimization”. *Computer Networks* 199 (2021).

[139] Jianzhong Shi et al. “Alarming situation of emerging H5 and H7 avian influenza and effective control strategies”. *Emerging microbes & infections* 12.1 (2023).

[140] Hyeong Jin Shin et al. “BERT-based spatial information extraction”. *Proceedings of the Third International Workshop on Spatial Language Understanding*. 2020, pp. 10–17.

[141] Marta S Shocket et al. “Transmission of West Nile and five other temperate mosquito-borne viruses peaks at temperatures between 23 C and 26 C”. *Elife* 9 (2020).

[142] Valérie Soti et al. “The potential for remote sensing and hydrologic modelling to assess the spatio-temporal dynamics of ponds in the Ferlo Region (Senegal)”. *Hydrology and Earth System Sciences* 14.8 (2010).

[143] Kim B Stevens, Marius Gilbert, and Dirk 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 (2013), pp. 1–14.

[144] Kim B Stevens and Dirk U Pfeiffer. “Spatial modelling of disease using data-and knowledge-driven approaches”. *Spatial and spatio-temporal epidemiology* 2.3 (2011), pp. 125–133.

[145] Thomas Stibor et al. “Geometrical insights into the dendritic cell algorithm”. *Proceedings of the 11th Annual conference on Genetic and evolutionary computation*. 2009.

[146] Corien Swaan et al. “Timeliness of notification systems for infectious diseases: A systematic literature review”. *PloS one* 13.6 (2018).

[147] Mehtab Alam Syed et al. “A Data-Driven Score Model to Assess Online News Articles in Event-Based Surveillance System”. *Information Management and Big Data: 8th Annual International Conference, SIMBig 2021, Virtual Event, December 1–3, 2021, Proceedings*. Springer. 2022, pp. 264–280.

[148] Mehtab Alam Syed et al. “A metadata approach to classify domain-specific documents for Event-based Surveillance Systems”. *2023 International Conference on Communication, Computing and Digital Systems (C-CODE)*. IEEE. 2023, pp. 1–5.

[149] Mehtab Alam Syed et al. *GeospaCy*. Version 1.0.0. 2023. DOI: 10.5281/zenodo.8415401. URL: https://github.com/mehtab-alam/GeospaCy/.

[150] Christine M Szablewski et al. “Reported global avian influenza detections among humans and animals during 2013-2022: comprehensive review and analysis of available surveillance data”. *JMIR Public Health and Surveillance* 9.1 (2023).

[151] Rachel A Taylor et al. “Predicting spread and effective control measures for African swine fever—Should we blame the boars?” *Transboundary and emerging diseases* 68.2 (2021), pp. 397–416.

[152] Rachel A Taylor et al. “The risk of infection by African swine fever virus in European swine through boar movement and legal trade of pigs and pig meat”. *Frontiers in veterinary Science* 6 (2020), p. 486.

[153] Jon Timmis et al. “An overview of artificial immune systems”. *Computation in cells and tissues: Perspectives and tools of thought* (2004), pp. 51–91.

[154] Annelise Tran et al. “Environmental predictors of West Nile fever risk in Europe”. *International journal of health geographics* 13 (2014), pp. 1–11.

[155] Sarah Valentin. “Extraction and combination of epidemiological information from informal sources for animal infectious diseases surveillance”. PhD thesis. Université Montpellier, 2020.

[156] Sarah Valentin, Renaud Lancelot, and Mathieu Roche. “Identifying associations between epidemiological entities in news data for animal disease surveillance”. *Artificial Intelligence in Agriculture* 5 (2021), pp. 163–174.

[157] Sarah Valentin et al. “Animal disease surveillance: How to represent textual data for classifying epidemiological information”. *Preventive Veterinary Medicine* 216 (2023).

[158] Sarah Valentin et al. “Dissemination of information in event-based surveillance, a case study of Avian Influenza”. *Plos one* 18.9 (2023).

[159] Sarah Valentin et al. “Elaboration of a new framework for fine-grained epidemiological annotation”. *Scientific Data* 9.1 (2022), p. 655.

[160] Sarah Valentin et al. “Monitoring online media reports for early detection of unknown diseases: Insight from a retrospective study of COVID-19 emergence”. *Transboundary and emerging diseases* 68.3 (2021), pp. 981–986.

[161] Sarah Valentin et al. “PADI-web: A multilingual event-based surveillance system for monitoring animal infectious diseases”. *Computers and Electronics in Agriculture* 169 (2020).

[162] A Viltrop et al. “African swine fever epidemiology, surveillance and control”. *Understanding and combatting African Swine Fever: A European perspective* (2021).

[163] Yanshan Wang et al. “Unsupervised machine learning for the discovery of latent disease clusters and patient subgroups using electronic health records”. *Journal of biomedical informatics* 102 (2020).

[164] Valdir Roberto Welte and Moisés Vargas Terán. “Emergency prevention system (empres) for transboundary animal and plant pests and diseases. the empres-livestock: an fao initiative”. *Annals of the New York Academy of Sciences* 10.1 (2004), pp. 19–31.

[165] World Organisation for Animal Health. *ASF Report*. Technical Report 33. Accessed: 2024-07-31. World Organisation for Animal Health, 2023. URL: https://www.woah.org/app/uploads/2023/05/asf-report33-002.pdf.

[166] Xinhua. *France detects bird flu at duck farm*. 2022. URL: https://english.news.cn/20220829/444b6757fa714034a9e6ea33a092bbc0/c.html.

[167] Mahendra Pal Yadav, Raj Kumar Singh, and Yashpal Singh Malik. “Emerging and transboundary animal viral diseases: Perspectives and preparedness”. *Emerging and transboundary animal viruses* (2020), pp. 1–25.

[168] Hua Yang et al. “A survey of artificial immune system based intrusion detection”. *The Scientific World Journal* (2014).

[169] Johanna J Young et al. “One Health approach for West Nile virus surveillance in the European Union: relevance of equine data for blood safety”. *Eurosurveillance* 24.16 (2019).

[170] Samira Yousefinaghani et al. “A decision support framework for prediction of avian influenza”. *Scientific Reports* 10.1 (2020), pp. 1–14.

[171] Victor L Yu and Lawrence C Madoff. “ProMED-mail: an early warning system for emerging diseases”. *Clinical infectious diseases* 39.2 (2004), pp. 227–232.

[172] W Zhou et al. “Earthquake prediction model based on danger theory in artificial immunity”. *Neural Network World* 30.4 (2020), p. 231.

[173] Wen Zhou and Yiwen Liang. “A new version of the deterministic dendritic cell algorithm based on numerical differential and immune response”. *Applied Soft Computing* 102 (2021).