EAI Endorsed Transactions
on Internet of Things
Research Article EAI.EU

Data prediction system in malaria control based on physio-chemical parameters of anopheles breeding sites

Kodzo M. Parkoo[1,*], Bamba Gueye[2] , Cheikh Sarr[1], and Ibrahima Dia[3]

[1] Université de Thiès, Thiès, Senegal
[2] Université Cheikh Anta Diop, Dakar, Senegal
[3] Institut Pasteur de Dakar, Dakar, Senegal

Abstract

Malaria is a public health problem in Senegal. As a result, a real program focused on prevention and treatment has been put in place to fight it. Despite the efforts made, the prevalence rate of malaria is still worrying. To have a prediction system that, once certain physicochemical information, will inform if we can or not attend to the development of anopheles larvae. Our work consisted of collecting data on mosquito breeding sites, processing, and analyzing them in order to predict the physicochemical conditions for the development of Anopheles larvae. Larval control is an alternative to reduce the prevalence rate of malaria. We retain logistic regression as an algorithm and water electrical conductivity, water turbidity, temperature, and dissolved oxygen as determinant parameters. The learning and prediction system set up on the basis of the determining parameters and logistic regression worked. The predictions will be improved by further training our system with field data.

Keywords: algorithms classification, larvae control, data analysis, data predictions, malaria, python.

Received on 15 August 2022, accepted on 30 November 2022, published on 15 December 2022

doi: 10.4108/eetiot.v8i4.2936

1. Introduction

According to the World Health Organization (WHO), in 2020 there were a total of 241 million cases in worldwide. The number of deaths due to malaria in the same year was 627,000. Africa bears a large part of these statistics. Thus, in 2020, 95% of cases come from Africa with 96% of deaths worldwide [1]. Malaria remains an endemic disease in Senegal, even though the number of malaria cases decreased by 38% between 2015 and 2019 (from 69 to 50 per 1,000 population) and the number of malaria deaths decreased by 7.1% during the same period (from 0.30 to 0.28 per 1,000 population) [2].

The main species found are *Plasmodium falciparum* (~ 98%) with rare cases of *P. ovale* and *P. malariae* throughout the country. The transmission of malaria parasites is ensured by seven anopheline species including four major vectors (*Anopheles gambiae* s.s, *An. coluzzii*, *An. arabiensis* and *An. funestus* s.s) and three secondary vectors of local importance (*An. melas*, *An. nili* and *An. pharoensis*). "An. gambiae s.s" represented by the molecular forms S and M (currently "*An. gambiae*" and "*An. coluzzii*" respectively) is present throughout the country but predominates in the humid southern areas while "*An. arabiensis*" is also represented in all regions and predominates in the drier central and northern areas. The species An. melas, is localized on the coast in the mangrove and along certain rivers up to the limits reached by the rise of marine salt water.

"*An. pharoensis*" is especially abundant in the lower valley of the Senegal River in the north, "*An. nili*" is most abundant in the southern regions, near the watercourses. In terms of management, the National Malaria Control Program of Senegal has adopted artemisinin-based combination therapy (ACT) as its first-line treatment and introduced rapid diagnostic tests (RDT) in 2007 [3]. In addition to treatment,

* Corresponding author. Email: Kodzo.mawuessenam@gmail.com

there is a prevention component with vector control and chemoprevention as the main focus. Chemoprevention is the use of drugs or drug combinations to prevent malaria infection and its consequences. Vector control includes two aspects: the use of insecticide-treated bed nets and indoor residual spraying. The proportion of households with insecticide-treated nets (ITNs) increased steadily from 2005 to 2017 (from 20% to 85%) but decreased in 2018 to 77 % [3].

Despite the decrease in the mortality rate from malaria, due of preventive measures, the situation remains critical with approximately 260 deaths in 2019 (Table 1).

Table 1. Level of morbidity and mortality indicators in 2018 and 2019 [3]

Year	Total of suspected cases	Total tested cases	Test realization Rate	Confirmed cases	Population 2019	Incidence per 1000 hbts	Malaria severe cases	Hospitalized per 10,000 hbts
2018	2,096,124	2,090;323	99.72%	530,944	15,663,116	33.9	13,350	9
2019	2,010,398	2,005,860	99.8%	354,708	16,209,119	21.9	9,352	5.8

Year	Total of deaths (all cases)	Malaria deaths cases	Proportional mortality	Rate of hospital lethality	Total malaria-related deaths for 10,000 hbts	Children under 5 years total death (all cases)	Malaria deaths cases (Children under 5 years)	Proportional mortality (Children under 5 years)
2018	15,745	555	3.5%	4.2%	4	4,575	147	3.2%
2019	15,623	260	1.7%	2.8%	1.6	4,871	62	1.3%

Year	Malaria deaths cases (Children under 5 years)	Malaria deaths cases	the under-five year mortality rate	Under-five population	Total malaria-related deaths for 10,000 children
2018	147	555	26.5%	2,950,931	5
2019	62	260	23.9%	3,079,733	2

One of the causes of this malaria situation is the resistance of mosquitoes to the preventive measures used in vector control. Indeed, there is an evolution and spread of resistance to insecticides, in particular to pyrethroids. This resistance threatens the effectiveness of the current standard vector control in many countries, namely long-lasting insecticide nets (LLINs) containing only pyrethroids [4].

It should be noted that all proposals for malaria control, have so far focused on the adult Anopheles. Therefore, it is imperative to adopt a focused approach of zone-specific interventions before and during larval development. Being interested in larval development implies studying the aquatic environment in which the larvae develop. The study of water is already done through quality monitoring systems based on the Internet of Things (IoT) [5]. We will therefore use a water quality analysis system to collect parameters related to the water surfaces that contribute to larval development. This approach, known as larval control (LC), is struggling to gain momentum in Senegal despite its inclusion in vector control [6]. This is due to the complexity of data collection and analysis involved in LC. Beyond data collection, which can be done by using the IoT, analysis requires the choice of algorithms as well as prediction. The choice of algorithms becomes crucial because if none are found that corresponds to the data to be processed, they must be built. In order to predict heart disease by analyzing a patient's retina, the American company Verily created a prediction algorithm based on Deep Learning [7]. Thus, thanks to prediction algorithms applied to data, we can participate in the LC by predicting the development or not of anopheles larvae. We will therefore present the IoT-based data acquisition architecture and the machine learning-based larval presence prediction mechanism. Our study makes, it possible to use water quality analysis systems in a different way, to have a database on the development of Anopheles larvae, to protect the environment by avoiding the systematic spraying of insecticides for mosquito control, and above all to reduce the rate of malaria contamination.

The rest of this paper is structured as follows: the second section deals with the different tools, methods used and implementations made in this research; section 3 presents the results of our implementations; section 4 proposes a discussion around our different results and finally the conclusion to gather our results and perspectives.

2. Materials and methods.

2.1 Identification of larvae

Mosquitoes are small insects (5 to 20 mm in size) with long, thin wings as shown in Figure 1. They differ from other Diptera such as flies by the presence of small scales on most of the veins and wings. Their body is slender, there are light brown to black, sometimes marked with spots and stripes [8].

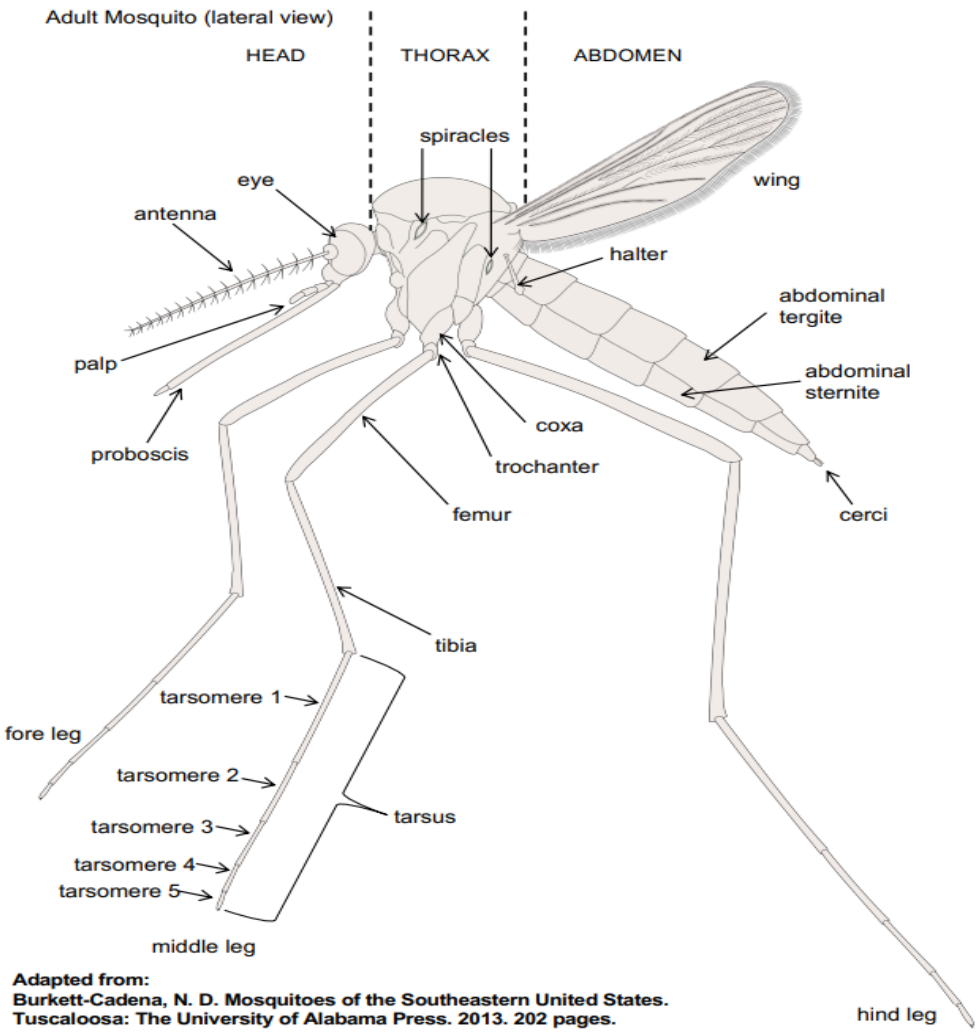

Figure 1. Morphology of the adult mosquito [9]

In the adult stage, the body is divided into three distinct parts: the head, the thorax, and the abdomen, each part having its own constituent parts. From eggs to adult mosquitoes undergo several metamorphoses (Figure 2).

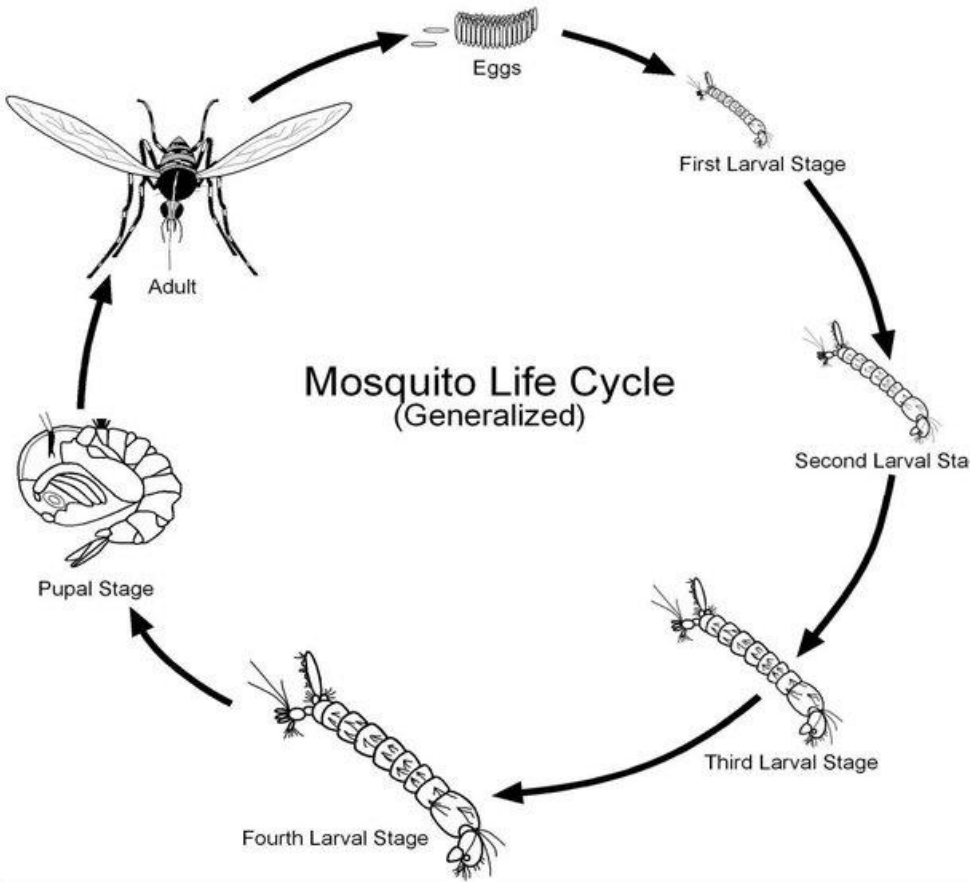

Figure 2. Representation of the life cycle of mosquitoes [10].

Mosquito breeding sites can be extremely diverse. For example, mosquito larvae can be found in permanent or temporary bodies of water, large or small, heavily polluted or clean; puddles, borrow pits puddles, borrow pits, silt, tires, and footprints. All mosquito species are fully metamorphosing or holometabolous insects, i.e., the different stages (egg, larva, pupa, adult) have different morphologies. Figure 2 illustrates the life cycle of mosquitoes. For instance, during their life cycle, they have an initial aquatic life (immature stages) then after the metamorphosis an aerial life (adults) [11]. Only the females are hematophagous. In most species, the female needs a blood meal to mature her eggs. The eggs develop into larvae and then into pupae. Emergence marks the transition from aquatic life to aerial life (Figure 2). After emergence, mating takes place, and then the females carry out the trophogonic cycle: search for a blood meal, rest for ovarian maturation of the eggs, and then oviposition [12].

As our study is based on larvae, we will focus aquatic life. By observation, we can distinguish larvae from pupae. Anopheles larvae can be recognized by the absence of legs and a relatively large thorax. Morphologically, the larva consists of three parts: the head, the thorax, and the abdomen. The head carries the 2 antennae, 2 large compound eyes, and a pair of mouth brushes which serve to create a water current bringing food particles to the mouth which is in the ventral position. The thorax is composed of 3 coalescent segments ensures the connection between the head and the abdomen. The latter includes 9 clearly visible segments, each with different ornamentation, including the tergal plate and accessory plates, setae, which may be simple or branched or used to recognize different species [8].

2.2 Data collection

The data we have exploited come from two different collection operations. The first data collected between 11/09/2018 to 12/15/2019, are based on a network of sensors. They allowed us to determine the reliable algorithms and also the factors influencing the presence of larvae. The data collected in the second stage, in October 2020, allowed us to learn and predict the presence of larvae of anopheles larvae.

2.2.1. First stage collection system

The measuring station represents the system used for data collection, show in Figure 3. He is constituted by an acquisition node, a gateway and a storage system.

Figure 3. Architecture of the measuring station constituted by an acquisition node, a gateway and a storage system.

2.2.1.1 Acquisition node

The acquisition system consists of a micro controller which all the sensors of the system are connected. It is powered by a solar power source (solar panel + battery) and is equipped with a LoRa module for transmitting data to the gateway.

Sensors

Sensors: In this system, four sensors are used for the measurement of water quality parameters (pH, Electrical Conductivity (EC), oxidation/reduction potential (ORP) and water temperature). However, the acquisition node is scalable to allow the addition of several sensors as needed. The Table 2 present the list of all the sensors used by acquisition node. The micro controller used to collect and process the parameters of the pool is an arduino Mega 2560 card [13].

Table 2. List of sensors [13]

Parameters	Sensors	Measurement range	Accuracy
Temperature	DS18B20	−10◦C to +85◦C	±0.5◦C
z pH	SEN0169	0 - 14	±0.1pH(25◦C)
ORP	SEN0165	−2000mV to 2000mV	±10mV (25◦C)
EC	DFR0300	1ms/cm-20ms/cm	≺ ±1ms/mm

Transmission Module

The pool is located about 500 meters from the building of the Department of Plant Biology where the Gateway is deployed. To ensure good communication, we used LoRa transmission. Since LoRa modulated signals are transmitted in ISM bands, we do not have to pay any fee to the local Telecommunication operator. Before deploying our system, as it was did in [14], we made coverage tests to choose an optimal position of the acquisition node to minimize the packet loss ratio. 624 packets were sent and we received 599. It gives us a packet error rate of 4% and the RSSI gateway is below −75 dBm. For the measurement station, the transmission between the sensor node and the gateway is done using structured frames showed in Figure 4 [13].

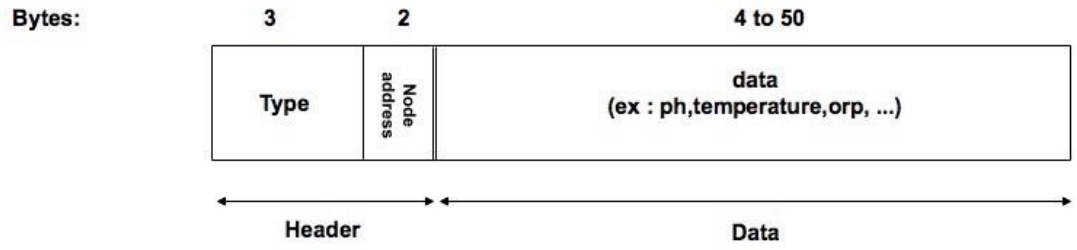

Figure 4. Frame [13]

Type is used to define frame type. There are two types of frame INF and CMD. INF frames are up link, data from sensors are encapsulated in these frames. CMD frames are down link in which commands are stored from platform to the acquisition node.

Node Address: this field is for destination node address.

Data: The payload is stored in it.

The type of the frame defines the size of the latter. For uplink frames (INF data frame) the size can reach up to 60 bytes and 15 bytes for down link frames (CMD type).

Flowchart of the Acquisition node

The acquisition node program is represented as a flowchart in Figure 5. We start the program with start function, the initialization function start. The LoRa Module is initialized with the basic configuration (spreading factor, coding rate, frequency), the basic parameters of the sensors are also initialized. Then in the loop we define a default waiting period of 10 minutes TIMEOUT

. After that delay, the data are collected from the sensors and encapsulated in a INF frame which is sent to the gateway through LoRa network. During the waiting delay, the node listens to the messages that would come from the server because it can receive a CMD frame at any moment. In a CMD frame, the server can whether ask for immediate data or can ask to the node to change the delay time. In those cases, the node collects the data from sensors and send it to the server and update the waiting TIMEOUT [13].

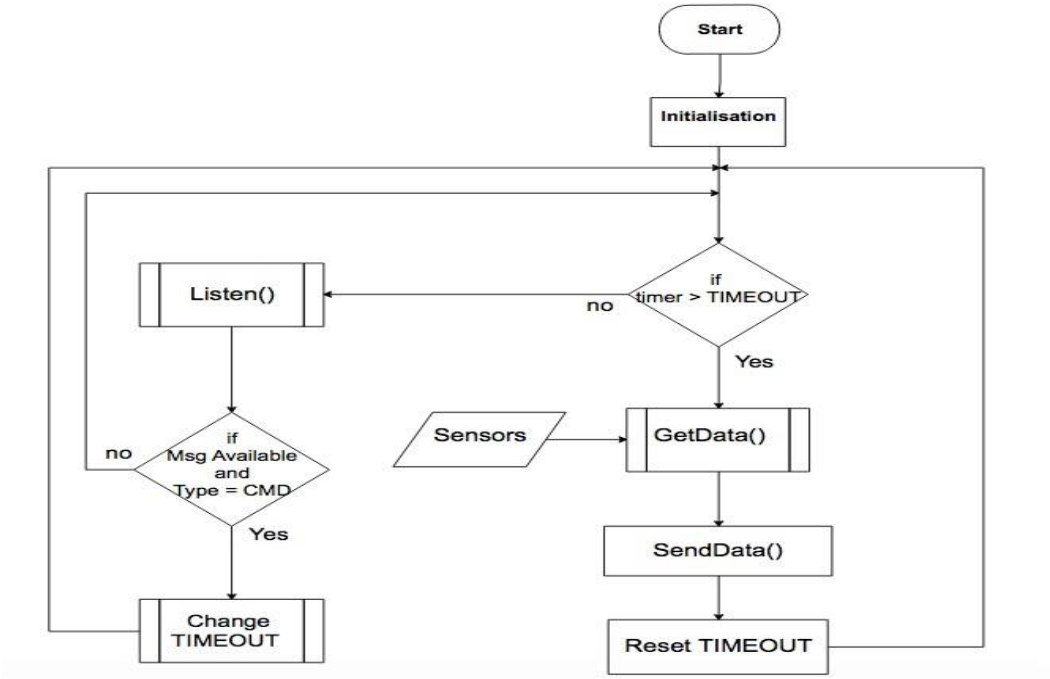

Figure 5. Acquisition node flowchart [13].

.

2.2.1.2 Gateway node

The gateway makes it possible to relay the frames between the acquisition node and the platform. It consists of a Dragino LG01 − P box which has four interfaces: LoRa,

WiFi, Ethernet, 3G/4G. Upon receipt of a frame from the LoRa network, the frame is sent to a local server via the Ethernet interface and stored in a database. After that, the data is then replicated to a cloud database if the internet connection is active. This replication enables to have a backup system and the possibility to access to collected data directly via Internet [13].

2.2.1.3 Data collection

The deployed measurement station is shown in figure 6 was installed at the UCAD botanical garden. The water reservoir on which the parameters were measured is not a controlled environment. All the data are grouped after the first processing, in a database of 1863 records.

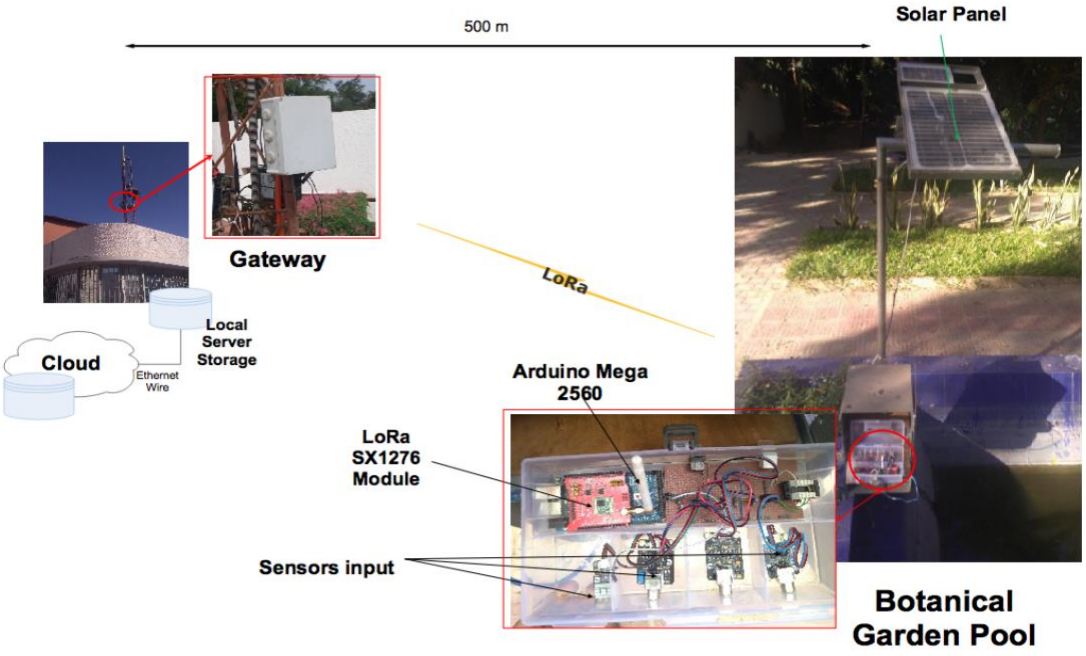

Figure 6. Measurement station system [13]

2.2.2 Second stage collection system

2.2.2.1 Sampling of aquatic stages and collection of physicochemical and biological parameters

Aquatic stages were collected in water collections in October 2020. After inspection at each site, larvae were collected either by dipping or pipetting method depending on the size of the sites and then placed in jars with the site number. For each site, the presence or absence of Culicidae other than Anopheles, vegetation, turbidity and sunlight were noted. After measuring the size of the gites (length, width, depth) with a decameter, the following parameters were then measured with a portable field tester (SD Card Real time Datalogger): temperature, amount of dissolved oxygen, salt content and pH.

The larvae were then sorted in the laboratory and stored in tubes containing 70° ethanol.

2.2.2.2 Study site

The proposed work was carried out in the district of Toubacouta (Senegal) and its surroundings due to the sympatric presence of the species "*An. arabiensis*", "*An. gambiae*", "*An. coluzzii*", and the observation of contrasting hybridization rates between the latter two. This area is located in the Sudanian domain between isohyets 700 and 1000 mm (Figure 7). The climate is Sudanian with a rainy season that extends over 4 to 5 months from June to September-October with temperature ranges often exceeding 25°C. Maximum temperatures are recorded between April-May (40°C) and October (35°C), while minimum temperatures often around 15°C are recorded in December-January and between 20°C and 25°C the rest of the year.

Figure 7. Different study sites

2.3 Study of some prediction algorithm

In the list of our data to be analyzed we can already distinguish numerical data (temperature, pH, width, and depth of the site...), count data (number of larvae identified in the water...), and binary data (yes, no...). We defined as predictive variable the values we know, the one that represents our input, and as response variable the one that will be used for the interpretation after the treatment. We proceeded to a summary presentation of some algorithms before testing them on our data.

2.3.1 Logistic regression

Logistic regression is a statistical model for studying the relationships between a set of qualitative variables Xi and a qualitative variable Y [15]. Its analysis model allows associating an event based on a qualitative variable and factors likely to influence this event (explanatory variables). In this particular case, the explanatory variables will be of the continuous numerical type (temperature, pH, width, and depth of a deposit) and the qualitative variable will be the presence (yes/no) of larvae. Therefore, it should be noted that the logistic regression does not directly provide the yes or no answer to the presence of larvae, as it is not these answers that are modeled. Rather, we model the probability of realizing one of these options. This probability is modeled by a sigmoid curve following the function:

$$f(x) = exp(x)/1 + exp(x) = p \quad (1)$$

As our function curve is bounded at 0 and 1, based on our continuous numerical explanatory variables, we have the possibility to determine the probability of the presence or absence of larvae.

Thus, we consider the following model:

$$pi = Prob(yi = 1\, j\, xi) = F(xib) \quad (2)$$

Depending on each variable, we can know whether it allows the development of anopheles larvae or not.

2.3.2. The random forest

Random forests are methods for obtaining predictive models for classification and regression. The method implements binary decision trees, in particular CART trees proposed by Breiman et al. (1984) [16]. It is a theoretical set composed by several decision trees. Each decision trees in the composition of the forest uses the same nodes but different data. It merges the decisions of several decision trees to find an answer, which is the average of all these decision trees. Our interest in this algorithm lies in the limitations of the decision tree.

Indeed, it happens that the generated trees are not balanced. This means that, when reading, one branch is longer than the others because of the results obtained. It is therefore recommended to balance the database before construction, to avoid one class dominating. Moreover, sometimes the generated trees are too complex and do not generalize well to the different cases, subject to processing. In addition, we note instability with some trees. This means that slight changes in the data produce very different trees and changes in the nodes near the root greatly affect the resulting tree [17].

2.3.3. Neural networks

A neural network, in its classical definition, is a computer system inspired by the human brain works to learn. If we go further, we will understand that it is a set of nodes linked by connections. Artificial neural networks (ANN) are organized in layers with an input layer, an invisible layer, and an output layer. Triggered by an input, the input nodes will trigger the other nodes by connections [18] [19]. In detail, it can be noted that direct links (or connections) between nodes are defined to know where the information goes (Figure 8). In addition, numbers (weights) are assigned to the connections so that some connections are stronger than others like real neurons.

.

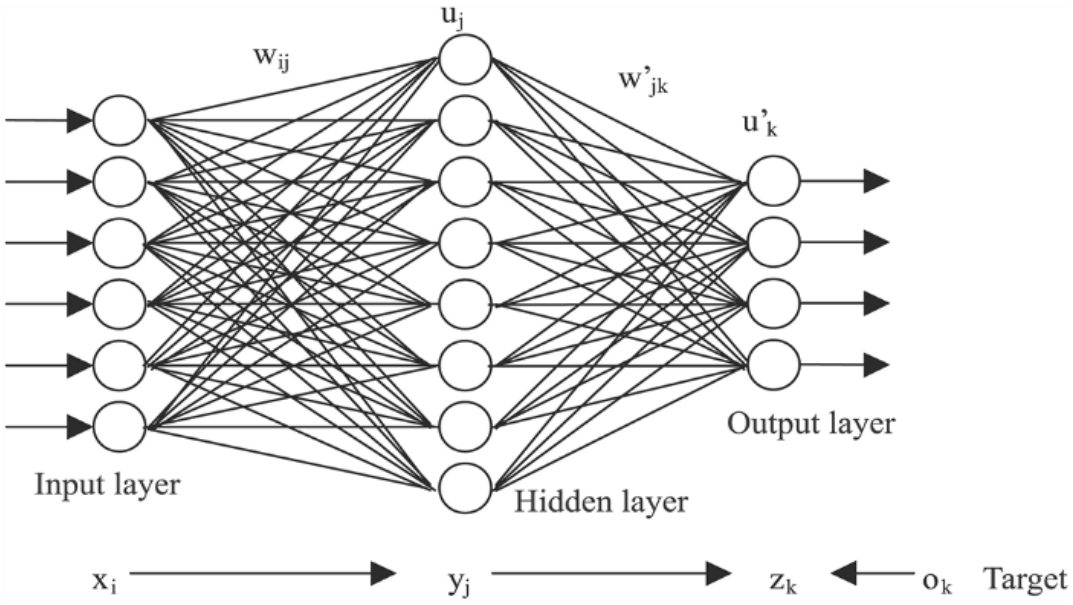

Figure 8. Simplified representation of a neural network [19].

2.3.4. Naive Bayes

Naïve Bayes, commonly used in machine learning, is a collection of classification algorithms based on Bayes theorem. It is not a single algorithm, but a family of algorithms. All these algorithms share a common principle, namely that, each classified feature is independent of the value of any other feature. Classifiers are based on the common principle that the value of a specific feature is independent of the value of any other feature. They allow us to predict the probability of an event occurring based on conditions we know for the events in question. The name comes from the Bayes theorem, which can be written mathematically as follows [20].

$$P(A|B) = P(B|A) * P(A)/P(B) \quad (3)$$

The application of the Bayes theorem to several variables makes the calculation complex. Thus, to get around this problem, one approach is to consider these variables independently of each other. This is a strong assumption. Generally, the predictor variables are interdependent.

2.4 Experimental settings

The implementation of our all algorithms was done in Python. To do this we deployed a virtual machine, Windows7, with 4 GB of RAM, 60 gigas of storage, and 4 Intel processors at 2.10 GHz. On this machine, we installed Anaconda 3 with the different applications: Jupyter Lab and Jupyter Notebook.

The main steps of the data processing were: the import of libraries, data normalization, and data exploration.

2.5 Algorithm reliability test

2.5.1. Presentation of data

2.5.1.1 Raw data

The data from our sensors, collected from our database in their raw state, are in Table 3. We have eighteen thousand six hundred and twenty-four (18624) lines of records.

Table 3. Presentation of raw data from sensors.

id,valeur,"sensor","date"
1,7,"ph","2017-07-02 04:04:33"
2,12,"orp","2017-07-02 04:04:33"
3,45,"electrical_conductivity","2017-07-02 04:04:33"
4,180.3,"turbidite","2017-07-02 04:04:33"
5,1.8,"temperature","2017-07-02 04:04:33"
6,1.7,"wind_speed","2017-07-02 04:04:33"
7,22.5,"direction","2017-07-02 04:04:33"
8,1.8,"humidity","2017-07-02 04:04:33"

2.5.1.2 Dataset cleaning and normalize

The presentation of our data as we have recovered them, does not correspond to the standard supported by Python. Indeed, the implementation of our various algorithms will be done with Python. So, we proceeded to a series of treatments in VBA in order to harmonize the data.

```
Sheets("REORDER").Select
 Sheets("REORDER").Cells.Select
  Selection.Delete Shift:=xlUp
  Sheets("REORDER1").Select
  Sheets("REORDER1").Cells.Select
  Selection.Delete Shift:=xlUp
  Sheets("DATARAW").Select
  Sheets("DATARAW").Columns("A:A").Select
  Selection.Copy
  Sheets("REORDER").Select
  Sheets("REORDER").Range("A1").Select
  ActiveSheet.Paste
```

```
''''''''''''''''''''''''''''''''''''''''''''REORDER THE DATA'''''''''''''''''''''''''''''''''''
Sheets("REORDER").Select
   Sheets("REORDER").Columns("A:A").Select
   Selection.TextToColumns Destination:=Range("A1"), DataType:=xlDelimited, _
       TextQualifier:=xlDoubleQuote, ConsecutiveDelimiter:=True, Tab:=True, _
       Semicolon:=True, Comma:=True, Space:=True, Other:=False, FieldInfo:= _
       Array(Array(1, 1), Array(2, 1), Array(3, 1), Array(4, 1)), TrailingMinusNumbers:= _
       True
```

Figure 9. Excerpt of VBA code for data ordering.

After executing the VBA code (figure 9), we obtain a new presentation of our data. In total we have one thousand eight hundred and sixty-three rows, as represented in the table 4.

The aim of the exploitation of these data is to allow us to find algorithms with a low error rate in the prediction and to find the most determining parameters for the presence of larvae. In addition, since we have about ten parameters to analyse, we try to retain analysis, and we try to retain only those parameters that really influence the presence of larvae.

Table 4. Presentation of data from the sensor network.

Id	Ph	ORP	C.E	Turb	W. Temp	Speed	Dir	Hum	Temp	Rain	Pres. Larvae
1	7	12	45	180.3	1.8	1.7	22.5	1.8	1.6	1.5	-
2	8	10	60	178.8	1.9	1.7	22.5	1.9	1.7	1.5	-
3	7	10	19	186.4	1.9	1.7	22.5	1.8	1.6	1.5	-

In Table 4, we are missing the observations related to the presence or absence of larvae. In order to make our data really exploitable, we will proceed to imputation the column "Presence of larvae" (Pres. Larvae). Imputation is the process used to assign replacement values to missing, invalid, or inconsistent values that have failed checks [21]. We, therefore, performed a "cold deck" imputation, which involves using an alternative source, such as historical data from a previous iteration. This allowed us to assign Boolean values (0 or 1) for the presence or non-presence of larvae.

Table 5. Presentation of data from the sensor network after imputation of missing data.

Id	Ph	ORP	C.E	Turb	W. Temp	Speed	Dir	Hum	Temp	Rain	Pres. Larvae
1	7	12	45	180.3	1.8	1.7	22.5	1.8	1.6	1.5	0
2	8	10	60	178.8	1.9	1.7	22.5	1.9	1.7	1.5	1
3	7	10	19	186.4	1.9	1.7	22.5	1.8	1.6	1.5	1

This assignment was done in the proportions of 50% for the presence and 50% for the non-presence (Table 5).

2.5.2. Performance evaluation

The goal is to determine which of our algorithms will be reliable. Thanks to the imputation of missing data, we already know the rate of presence and non-presence of larvae. Therefore, we divided our dataset into two parts with a rate of 60 percent and 40 percent. The first part of the 60 percent was used for training by our algorithms and the second part of the 40 percent, for prediction. Knowing that we have a rate of 50 percent for each case, thanks to the imputation of data, we can evaluate the reliability of each algorithm according to its prediction rate. The more reliable the algorithm, the higher its prediction rate.

The different algorithms are assigned to models for reliability testing. Each algorithm will generate its result.

Through the reliability test of the algorithms, three of them, namely logistic regression, linear discriminatory analysis, and random forest, are in the lead. On the particularities, only the logistic regression allows to establish the correlation between a parameter and the presence of larvae. This is due to the mode of representation of the variable "Presence of larvae" which is binary. It should be remembered that logistic regression is a classification algorithm that allows two classes to be determined (true/false or yes/no). The objective of our analysis at this stage is to determine whether a parameter is a determining factor for the presence of larvae or not. Thus, we will establish two classes, namely determining and non-determining parameters.

Based on the distributions obtained by imputing missing data, we already know the rate of larvae present on the whole data set. Nevertheless, we have evaluated the presence rate on the 60 percent of data intended for training (Figure 10).

```
# Percentage of larvae present
count_lav = len(data_frame[data_frame['presence_larves']==1])
count_no_lav = len(data_frame[data_frame['presence_larves']==0])
pres_of_lav = count_lav/(count_lav+count_no_lav)
print("Percentage of larvae present", pres_of_lav*100)
```

```
Percentage of larvae present 49.40923737916219
```

Figure 10. Assessment of the percentage of presence of larvae.

2.5.3 Justification of the choice of the algorithm

In the result section we have the algorithm reliability test result, which help for the algorithm selection. In addition, there are reasons for the choice of logistic regression. Logistic regression is a method of statistical analysis that predicts a binary result, for example yes or no, from earlier observations of a dataset. It's predicting a dependent data variable by analyzing the relationship between one or more independent variables in existence. This corresponds to our study case.

2.6 Search for parameters determining the presence of larvae

The value of larvae present percentage (Figure 10), will be used as a control in the evaluation of the parameters.

Logistic regression, a classification algorithm, is implemented in python to graphically represent the correlation between our different parameters (Ph, Orp, Electrical Conductivity, Turbidity, temperature, wind speed, wind direction, humidity, ambient temperature, rain) and the presence of larvae.

```
table= pd.crosstab(data_frame.ph,data_frame.presence_larves)
table.div(table.sum(1).astype(float), axis=0).plot(kind='bar', stacked=True)
plt.title('Ph / Larvae')
plt.xlabel('Ph')
plt.ylabel('PH Influence')
```

Figure 11. Creation of graph evaluation in relation to the presence of larvae.

The graphs that will result from the code in Figure 11 will allow us to differentiate the parameters that determine the presence of larvae.

The starting assumption is that the rate of presence of larvae rate is 49.409 percent, so all the parameters that will cause a variation of more than 50 percent are therefore determining for the presence of larvae.

2.7 Learning and prediction of the presence of larvae by logistic regression.

Presentation of data.

The data we use are from the second stage of collection. These data were collected in October 2020 by a team from the Institut Pasteur of Dakar in Toubacouta. These initial data include information on the breeding sites, the chemical characteristics of the breeding sites, and the presence or absence of larvae (Table 6). Since for our study only the physico-chemical characteristics and the presence of larvae are of interest, our final data source will retain only the latter.

Table 6. Presentation of data from the second collection.

PH	Temp	Conductivity	Salinity	Dissolved oxygen	Turbidity	Sunshine	Vegetation	Status
7.45	32.5	43.9	0	11	0	0	0	1
7.68	34.1	227	0	8.7	1	0	1	0
7.26	30.2	52.6	0	20.2	1	0	0	1

With python, we will keep the chemical parameters, some physical, and the presence or absence of larvae represented by «Status» in Table 6.

Logistic regression use

The objective is to train the system to predict the presence or absence of larvae. This training will be done with logistic regression in Python. We proceeded to segment our data set in the proportions of 50 percent for training and 50 percent for prediction. Knowing that "Vegetation" and "Sunshine" are not determinants for the presence of larvae we obtain the data structure of Figure 12.

```
[3]  dataset = pd.read_excel('Fichier_larves_INH_2_N.xlsx')

[4]  dataset.shape

     (4794, 9)

[5]  dataset.drop(['Ensoleillement'],axis='columns',inplace=True)
     dataset.drop(['Vegetation'],axis='columns',inplace=True)
     dataset.head()
```

	pH	Temperature	Conductivite	Salinite	Oxygene_dissout	Turbidite	Statut
0	7.45	32.5	43.9	0.0	11.0	0	1
1	7.68	34.1	227.0	0.0	8.7	1	0
2	7.26	30.2	52.6	0.0	20.2	1	1
3	6.87	32.7	25.5	0.0	5.2	1	1
4	7.70	33.6	68.9	0.0	23.6	1	1

Figure 12. Presentation of data for prediction.

In order to evaluate the predictive quality of our model, we used the confusion matrix. The confusion matrix is like a summary of the estimation results for a particular classification problem. The confusion matrix, also known as the contingency table, is used to evaluate the performance of a classification model. In its simplest form, it compares the actual data for a target variable with that predicted by the model [22]. Figure 13 depicts the confusion matrix for our data.

```
nt, pf, nf, pt = cm.ravel()
```

```
print('Number of positive true : ', pt)
print('Number of negative true : ', nt)
print('Number of positive false : ', pf)
print('Nombre of negative false : ', nf)
```

```
Number of positive true :  2193
Number of negative true :  0
Number of positive false :  204
Nombre of negative false :  0
```

Figure 13. Confusion matrix.

Having an idea of the predicted versus actual results, thanks to the confusion matrix, we can proceed with our prediction.

3. Results

3.1 Algorithm reliability testing.

On the reliability test of the algorithms, it appears that not all of them are reliable in predicting the types of data we process. Through the above-mentioned process, we obtain the following reliability rates in percentage (Table 7).

Table 7. Ranking of algorithms by reliability rate.

Algorithm name	Prediction Rate of reliability
Logistics Regression	56.84
ALD	56.84
Random Forest	52.02
Naive Bayes	52.01
KNN	51.71
Neural Networks	50.04
SVM	42.90

3.2 Search for parameters determining the presence of larvae

Each parameter was studied in its intervention on a positive or negative result on the presence of larvae. The presentation of the result is done graphically. A parameter that influences the presence of larvae should be marked as either 0 or 1 on the graph (Figures 14 and 15).

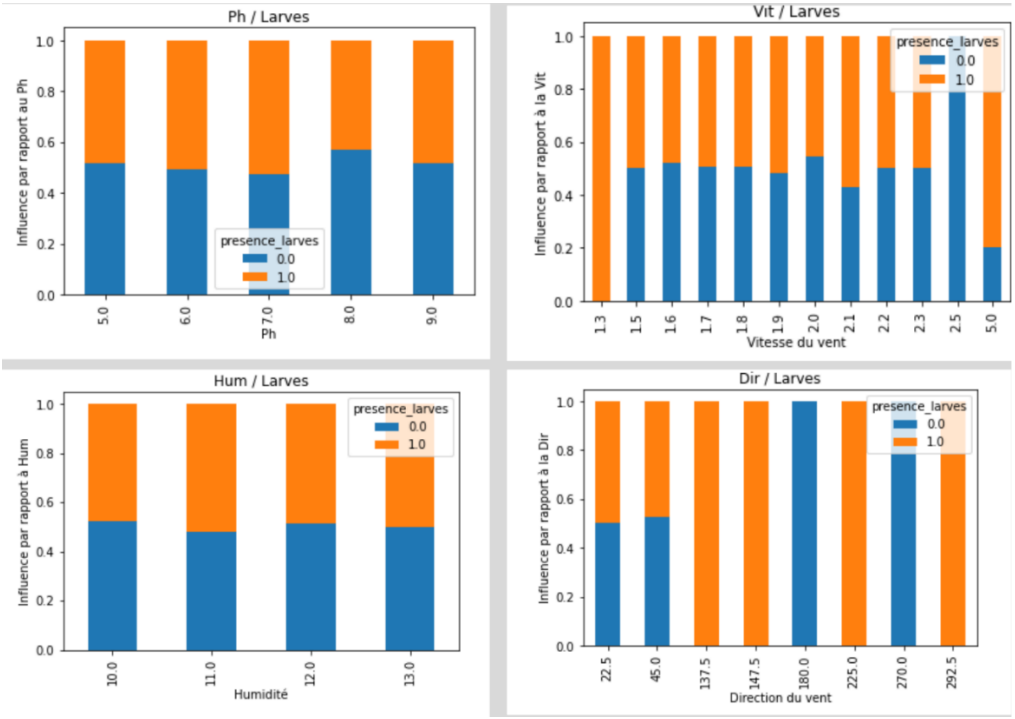

Figure 14. Graphs of the non- determining parameters.

.

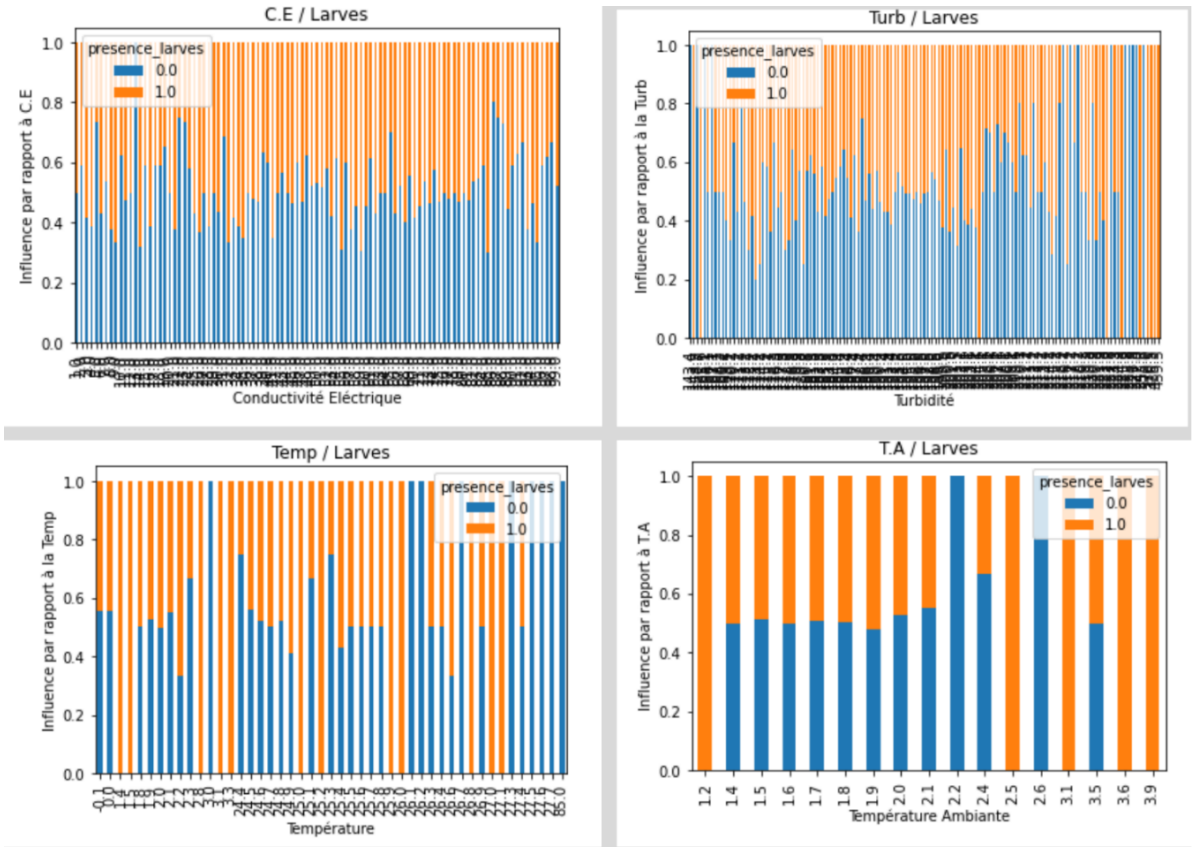

Figure 15. Graphs of the determining parameters.

As all parameters are treated separately, Figure 14 represents those with few peaks at 0 or 1. Figure 15, on the other hand, looks like parameters with numerous peaks at 0 or 1. By interpretation we can distinguish between parameters that are decisive for the presence of larvae and those that are not.

3.3 Data learning and prediction

Based on the learning performed by our system, each time a sequence of values corresponding to the determining parameters is submitted to it, it predicts whether these parameters lead to the presence or not (1 or 0) of anopheles larvae. Figure 16 is an example of parameter insertion for prediction.

```
[15] x_predict = np.array([[7.45,32.5,43.9,0,11,0]], dtype=float)
     classifier.predict(x_predict)

     array([1])
```

Figure 16. Insertion of values for prediction

4 Discussions

Regarding the results of the reliability tests of the algorithms, we have Logistic Regression and Linear Discriminant Analysis with 56.84 percent, Random Forest with 52.02 percent, Naives Bayes with 52.01 percent, and Neural Networks with 50.04 percent. Neural Networks with 50.04 percent. Our logic established through the imputation of missing data, allowed us to retain the algorithms with a high percentage. Thus, we can choose Logistic Regression and Linear Discriminant Analysis. Except that in our particular case, which aims to classify data into two categories: those of the presence of larvae or not, the Logistic Regression was chosen because it is a classification algorithm. Logistic regression, once chosen, allowed us to determine the parameters that have an influence on the presence of larvae.

A parameter that influences the presence of larvae must point to either to 0 or 1 on the graph. Otherwise, its values have no effect on the presence or absence of larvae. The distinction of the determining parameters was important because the deployed sensor network consisted of several sensors. Not knowing the determining factors meant, that all sensors had to be maintained, repaired and changed as necessary, at great cost. Having obtained the determining factors, we were able to deploy only those sensors. In addition, we were able to reduce the number of parameters to be analyzed in the prediction. As a result, we can guarantee a moderate use of resources for the prediction. This was also useful when retrieving data from the second collection. Although pH does not directly influence the presence of larvae, it does influence the other determining parameters. This explains its presence in the determinant parameters. So, only the determining parameters were used for the prediction.

Speaking of prediction, this was done with data from the second collection, which had the particularity of indicating the presence or absence of larvae. Thus, the learning was done directly on data strongly correlated to the presence of Anopheles larvae. This allowed us to elaborate a confusion matrix between the real and predicted facts. In order to understand our confusion matrix, let us remember that we have two (02) possible outcomes: the presence of larvae (positive result) and the absence of larvae (negative result). In our confusion matrix, these two (02) basic results are identified as "Positive True " (PT) and "Negative True" (NT). The remark we make is that the NT is 0, which reveals that in our data source we have very few records leading to the absence of larvae. That is to say that the parameters have been measured on sites where most of the larvae were present. This had an impact on the prediction as we obtained 204 " Positive False" (PF). This means that we predicted the presence of larvae when the actual parameters predicted the absence of larvae. However, this can be explained by the high presence of larvae in the records of our data source. In the end, we succeeded in setting up our prediction system, which manages to learn on the basis of parameters determining the presence of larvae. Moreover, this system allows us to make the prediction even if it requires adjustments. This ability to predict the presence of larvae in the African context is a major advantage. According to Centers for Disease Control and Prevention (CDC), it is difficult, if not impossible, to predict when and where the breeding sites will form, and to find and treat them before the adults emerge. Therefore, larval mosquito control for the malaria prevention in Africa has not been attempted on a large scale [23]. Successfully piloting

and deploying this prediction system in Senegal will be an opportunity for Africa. But to reach that point, we need to improve our prediction system and make it accessible to all, in order to perfect it.

5 Conclusion

Larvae control is a solution to reduce the prevalence of malaria by all means. Thus, the study of the algorithms and the determination of the parameters determining the presence of larvae brought us closer to prediction. We, therefore, retain logistic regression as an algorithm and parameters such as the pH, the electrical conductivity of water, water turbidity, temperature, and dissolved oxygen as a determinant.

The learning and prediction system implemented on the basis of the determinants parameters and logistic regression also worked. However, the latter has a deficiency related to the data it uses for its learning. Indeed, through to the confusion matrix, we understood that the system learned to predict the presence of larvae better than its absence. This imperfection can be corrected by training the system even better on other data.

In future work, based on the results obtained, we will improve the predictions, training our system even more. This will involve searching and processing new data that will be made available to our system.

References

[1] World Health Organization, *World malaria report 2021*. Geneva: World Health Organization, 2021. Consulted the: 12 May 2022. [Online]. Available on: https://apps.who.int/iris/handle/10665/350147

[2] World Health Organization, *World malaria report 2020: 20 years of global progress and challenges*. Geneva: World Health Organization, 2020. Consulted the: 12 May 2022. [Online]. Available on: https://apps.who.int/iris/handle/10665/337660

[3] U.S. President's Malaria Initiative Senegal Malaria Operational Plan FY 2020. Retrieved from (www.pmi.gov). Consulted the: 13 may 2022. [Online]. Available on: https://d1u4sg1s9ptc4z.cloudfront.net/uploads/2021/03/fy-2020-senegal-malaria-operational-plan.pdf

[4] World Health Organization, *Insecticide-treated nets for malaria transmission control in areas with insecticide-resistant mosquito populations: preferred product characteristics*. Geneva: World Health Organization, 2021. Consulted the: 17 May 2022. [[Online]. Available on: https://apps.who.int/iris/handle/10665/339542

[5] J. Mabrouki, M. Azrour, et S. E. Hajjaji, « Use of internet of things for monitoring and evaluating water's quality: a comparative study », *Int. J. Cloud Comput.*, vol. 10, nº 5/6, p. 633, 2021, doi: 10.1504/IJCC.2021.120399.

[6] Programme National de Lutte Contre Le Paludisme (PNLP), Plan Strategique National De Lutte Contre Le Paludisme Au Senegal 2021 - 2025. Consulted the: 12 may 2022. [Online]. Available on: https://senegal-cocreation.com/wp-content/uploads/2021/02/PSN_PNLP_Senegal_Version-finale_-Fevrier-

[7] R. Poplin *et al.*, « Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning », *Nat. Biomed. Eng.*, vol. 2, nº 3, p. 158-164, March 2018, doi: 10.1038/s41551-018-0195-0.

[8] P. Carnevale et V. Robert, Éd., « 2. Morphologie », in *Les anophèles : Biologie, transmission du Plasmodium et lutte antivectorielle*, Marseille: IRD Éditions, 2017, p. 22-46. Consulted the: 21 May 2022. [Online]. Available on: http://books.openedition.org/irdeditions/10388

[9] D. N. Burkett-Cadena, « Morphology of Adult and Larval Mosquitoes », p. 14.

[10] E. Rogozi, « MOSQUITO TRAPPING IN RECREATIONAL PARKS OF SELANGOR AND THEIR ROLE IN PUBLIC HEALTH », 2010, doi: 10.13140/2.1.4697.7608.

[11] N. Becker, Éd., *Mosquitoes and their control*, 2nd ed. Heidelberg: Springer, 2010.

[12] A. N. Clements et A. N. Clements, *Development, nutrition and reproduction*, Print on demand ed. Wallingford: CABI Publ, 2008.

[13] B. Ngom, M. Diallo, B. Gueye, et N. Marilleau, « LoRa-based Measurement Station for Water Quality Monitoring: Case of Botanical Garden Pool », in *2019 IEEE Sensors Applications Symposium (SAS)*, Sophia Antipolis, France, mars 2019, p. 1-4. doi: 10.1109/SAS.2019.8705986.

[14] M. R. Seye, B. Ngom, B. Gueye, et M. Diallo, « A Study of LoRa Coverage: Range Evaluation and Channel Attenuation Model », in *2018 1st International Conference on Smart Cities and Communities (SCCIC)*, Ouagadougou, juill. 2018, p. 1-4. doi: 10.1109/SCCIC.2018.8584548.

[15] « La régression logistique, qu'est-ce que c'est ? », *Formation Data Science | DataScientest.com*, 4 novembre 2020. https://datascientest.com/regression-logistique-quest-ce-que-cest (Consulted the 19 May 2022).

[16] « Forêts aléatoires de classification et de régression », *XLSTAT, Your data analysis solution*. https://www.xlstat.com/fr/solutions/fonctionnalites/forets-aleatoires-de-classification-et-de-regression (Consulted the 20 May 2022).

[17] L. Rokach et O. Maimon, *Data Mining with Decision Trees: Theory and Applications*, 2e éd., vol. 81. WORLD SCIENTIFIC, 2014. doi: 10.1142/9097.

[18] +Bastien L, « Réseau de neurones artificiels : qu'est-ce que c'est et à quoi ça sert ? », *LeBigData.fr*, 5 avril 2019. https://www.lebigdata.fr/reseau-de-neurones-artificiels-definition (Consulted the 20 May 2022).

[19] W. G. Baxt, « Use of an Artificial Neural Network for Data Analysis in Clinical Decision-Making: The

Diagnosis of Acute Coronary Occlusion », *Neural Comput.*, vol. 2, nº 4, p. 480-489, déc. 1990, doi: 10.1162/neco.1990.2.4.480.

[20] « Les Algorithmes de Naïves Bayes », *Analytics & Insights*, 1 March 2019. https://analyticsinsights.io/les-algorithmes-de-naives-bayes/ (Consulted the 20 May 2022).

[21] S. C. Government of Canada, « 3.4.4 Imputation », 2 septembre 2021. https://www150.statcan.gc.ca/n1/edu/power-pouvoir/ch3/imputation/5214784-eng.htm (Consulted the 7 August 2022).

[22] « Comment lire et exploiter une matrice de confusion ? », *Formation Data Science | DataScientest.com*, 16 February 2021. https://datascientest.com/matrice-de-confusion (Consulted the May 2022).

[23] C.-C. for D. C. and Prevention, « CDC - Malaria - Malaria Worldwide - How Can Malaria Cases and Deaths Be Reduced? - Larval Control and Other Vector Control Interventions », 16 July 2020. https://www.cdc.gov/malaria/malaria_worldwide/reduction/vector_control.html (Consulted the 5 October 2022).