International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
IJISAE
ISSN:2147-6799 www.ijisae.org Original Research Paper

Multilabel Classification for Predicting Crop Pests in Niger

Mahaman Lawali Inoussa Garba*[1], Harouna Naroua[2], Chaibou Kadri[3], Madougou Garba[4], Maman Aminou Ali[5]

Submitted: 15/01/2025 **Revised:** 27/02/2025 **Accepted:** 20/03/2025

Abstract*:* Crop pests pose serious threats to agricultural production and food security. With the advent of climate change in Niger, pest attacks have become increasingly frequent. This has become a crucial problem and a priority for farmers and government, as it can destroy the crop or harvest, thereby causing economic harm to the detriment of farmers and the population. Machine learning techniques are widely used in crop pests' prediction. However, the existing approaches generally focus on the prediction of crop pests using traditional classification methods. These approaches are limited, as they do not make it possible to predict multiple crop pests. Thus, simultaneous and rapid prediction of multiple pests remains a major challenge. In this study, we proposed an approach to predict all the pests of a crop in various localities by using multilabel classification techniques. We developed and compared nine (9) multilabel classification models over two different periods (monthly and annual) using historical data on crop pest infestation and climate. The classifiers are evaluated using Hamming Loss (HL). It was observed that the Radom k-labELsets (RAkEL) classifier is better both on monthly and annual prediction of all pests, with a comparative HL percentage value of 3.63% and 5.1%, respectively. This study extends the models available for crop pest prediction and opens a new path to improving the prediction of crop pests.

Keywords: *Machine Learning; Multilabel Classification; Prediction; Crop pest; Agriculture*

1. Introduction

Predictive modelling is a field of machine learning that allows learning from data (known observations) to predict a target variable (unknown in advance). This target variable is usually a single continuous or discrete variable, corresponding to common regression and classification tasks, respectively [1]. However, in practically relevant problems, there are several properties of interest which are target variables. These problems include annotating images with multiple labels, predicting gene functions and drug effects [1]. Problems with multiple binary variables as targets correspond to supervised learning problems where an instance can be associated with several predefined labels (target variables) belonging to multilabel classification task [1]. Recently, the issue of multilabel classification has drained considerable interest from the machine learning community, driven by a growing number of new applications in several broad and diverse domains, including text, audio, images, video, bioinformatics, and references [1, 2]. In this study, we apply it in the domain of plant protection.

With an area of 1,267,000 km2, Niger is one of the largest countries in Africa. In this country, agriculture is the most important sector of the economy. It represents more than 40% of the national gross domestic product (GDP) and constitutes the main source of income for more than 80% of the population [3]. The performance of the agricultural sector is nevertheless very unstable due to its high exposure to climate change.

1 Department of Mathematics and Computer Science, Faculty of Science and Technology, Abdou Moumouni University (UAM), Niamey, Niger

ORCID ID : 0009-0006-6899-5464

2 Department of Mathematics and Computer Science, Faculty of Science and Technology, Abdou Moumouni University (UAM), Niamey, Niger

ORCID ID : 0009-0003-4969-7224

3 Department of Mathematics and Computer Science, Faculty of Science and Technology, Abdou Moumouni University (UAM), Niamey, Niger

ORCID ID : 0009-0005-3528-0633

4 Directorate General of Plant Protection (DGPV), Ministry of Agriculture, Niamey, Niger

ORCID ID : 0009-0002-5208-2215

5Fermer Federation FUMA Gaskiya, Maradi, Niger

ORCID ID : 0009-0003-2735-1038

** Corresponding Author Email: mahamanlawali09@gmail.com*

Pests constitute a serious threat to food security and a major concern for farmers as they directly affect the yield and consequently decrease agricultural production. Accurate prediction of pests, for early and necessary treatment measures, is very difficult and requires considerable experience and expertise. Conventional pest identification techniques (visual inspection of plants, farmer's experience, personal mind and intuition) rely on human intervention. Poor judgment or delay in the identification and decision making process, can have detrimental impacts on production [4].

Nowadays, many computer-aided systems are used in almost all countries. They apply modern approaches such as machine learning and deep learning to increase the rate of pest recognition and the accuracy of results. However, most of the works were focused on the identification of the pest at the start or after the attack on the crop and thereby proposing necessary measures [5]. Other studies focused on predicting the occurrence of a single crop pest at a time, without any possibility of predicting a set of pests likely to attack the crop [6, 7]. Existing systems have considerable limitations such as late identification of pests (since the crop must first be attacked before identifying the pest) and the inability to predict all the pests of a crop. Moreover, although some studies have investigated the use of machine learning algorithms to predict crop pest in other countries/regions, there are gaps in understanding their effectiveness for the context of Niger.

This study aims to fill the gap in the literature by proposing a model (using multilabel classification and climate parameters) to predict crop pests over a period (month or year) in a locality, in order to provide farmers with enough time to plan effectives actions in time, thereby protecting their crops and production. For a given period (month, year), a crop can be infested by several pests. Therefore, pest prediction problem is a multilabel classification problem, since it involves predicting all pests of a crop. In this work, we will take a very practical approach to building a crop pest prediction model using the concept of multilabel classification. Two data sets are used. The first consists of real information on crop infestations by pests collected from the Directorate General of Plant Protection (DGPV) of Niamey. The second concerns climate data collected on the NASA web site. The aim is to provide farmers with a tool that allows them to secure their production by providing early protection from pest attacks.

2. Related work

In this section, we review some important works done in the area of machine learning applications for pest detection and diagnosis.

The study in [7] focused on proposing a predictive model based on deep learning to prevent diseases and pest infestations using environmental crop growth data. The data used for training the model was collected from the public database called “plant disease-causing dataset” provided by AI-hub, and for model validation internal data on strawberry gray mold were collected from November 2020 to May 2021. They considered five crops, two pest infestations per crop, and factors such as air temperature, relative humidity, dew point and CO2 concentration. The performance of the model was evaluated using AUROC and the results obtained show high performance with an average AUROC of 0.917.

A model was developed by [5] for plant disease prediction based on thermal images and using deep learning techniques. The system consists of a convolutional neural network (CNN) composed of three convolutional layers to overcome the computational overhead and overfitting problem for small datasets. The used pre-processed dataset consists of 1,044 thermal and visual images of selected leaves of 288 plants captured using a FLIR C2 camera and includes four classes (Normal, Stage1, Stage2, and Stage3). The performance of the model was evaluated using accuracy, precision, type I error and type II error, and was compared with four standard deep learning models (VGG-16, VGG-19, Resnet50 and Resnet101) and two machine learning algorithms (Linear Regression and SVM). The result shows that their model is better with an accuracy of 95%, a precision of 97.5%, a type I error of 2.3% and a type II error of 7.7%.

In [8] a deep learning model was proposed for tea smut disease classification using RGB and hyperspectral images. The dataset consists of sample images of tea plants from several plots of the Chunxi tea garden collected randomly over two seasons (250 samples in spring and 400 samples in autumn). The RGB images consist of 700 photos captured using a digital camera in natural light conditions, and the hyperspectral images constituting a total spectral matrix of 650 × 176, were acquired using a hyperspectral camera in a constructed cubic dark box. After preprocessing the data, they implemented various models, including ResNet18, VGG16, AlexNet, WT-ResNet18, WT-VGG16 and WT- AlexNet based on RGB images, and UVE-LSTM, CARS-LSTM, NONELSTM, UVE-SVM, CARS-SVM and NONE-SVM based on hyperspectral images. The performances of the models were compared using accuracy. The results showed that the WT-ResNet18 model is better for RGB with an accuracy of 70% and the CARS-LSTM model is better for hyperspectral with an accuracy of 95% higher than RGB.

In his study, [9] proposed an optimized CNN model using a genetic algorithm for the classification of pest types. Three datasets were used, including the Deng dataset consisting of 563 images and 10 classes, the Xie2 dataset named D0 composed of 4508 images and 40 classes and the Wu dataset named IP102 consisting of 75222 images and 102 classes. The performance of the model was compared based on accuracy with three CCN models at different scales, namely MobileNetV2, DenseNet121, and InceptionResNetV2. The results show that the optimized model offers performance closer to the literature with an accuracy of 99.89% on the D0 dataset, 97.58% on the Deng dataset and 71.84% on the IP102 dataset.

In [6], a system was proposed to predict plant disease using the Social Internet of Things (Social IoT) and deep learning techniques. They used three different datasets namely: the PlantVillage dataset for training the models; the Coffee Leaf Rust dataset containing data on coffee plants sampled by sensors on average 7 times per day for three months, and including attributes such as ambient humidity and temperature, pH, soil moisture, soil temperature and illuminance; and a dataset consisting of photos of plants. The last two datasets were used for model evaluation. They implemented and evaluated the performance of four different CNN architectures (DenseNet121, MobileNet, MobileNetV2 and NasNetMobile) based on accuracy and F1-score. The results show that MobileNetV2 is better with an accuracy of 94.58% and an F1-score of 94.58%.

In [10], random forests were used to create a model capable of identifying healthy and diseased papaya leaves. The model was trained on 160 images of papaya leaves taken on a plain background to eliminate occlusion. The proposed work involves various implementation phases: creation of the dataset, feature extraction, training of the classifier, and classification. The model was compared with other machine learning models such as SVM, K-nearest neighbors, naive Bayes, logistic regression, and CART based on accuracy. The results show that the proposed model outperforms the other techniques, with an accuracy of 70.14%.

In [11], a rice plant disease detection system was proposed using machine learning approaches. They were interested in the three best-known rice diseases, namely leaf smut, bacterial leaf blight, and brown spot diseases. The system takes images of rice leaves as input and applies different machine learning techniques such as k-nearest neighbors, decision trees, naive Bayes, and logistic regression to predict leaf diseases with different degrees of accuracy. The data used includes 480 instances, of which 432 (90%) instances were used for training and 48 (10%) instances for testing. After training and testing the models, a comparative study showed the supremacy of the decision tree model, with an accuracy of 97.9167%.

In [12], a real-time method was presented for corn leaf disease recognition using a convolutional neural network. The system works offline on a mobile device. The model is first trained with a large amount of suitable data and tested on a computer before being embedded to mobile devices. The role of the mobile device is to capture images of plant leaves using the camera, preprocess them, and pass them to the system for diagnosis. The model accuracy is up to 88.46%.

In [13], a cassava disease identification system was proposed using a convolutional neural network. Five categories of cassava leaf diseases were used with 10,000 images (training data) labelled and collected during a survey in Uganda. Given that the data is unbalanced, they combined the class weight, SMOTE (Synthetic Minority Over-sampling Technique), and focal loss techniques with the convolutional neural network to obtain an accuracy of over 93%.

In [14], a method to identify diseases in tomato crops based on leaf image analysis was introduced. The convolutional neural network approach was used for disease detection and classification. The data used was downloaded from plant village and includes nine disease classes and one healthy crop class. They used 10,000 images for training, 7,000 images for validation, and 500 images for testing. Since the number of images within classes is not balanced, they applied data augmentation techniques to balance the images within classes. Their technique offers a higher accuracy of 91.2% compared to models such as Mobilenet, VGG 16, and InceptionV3 with 63.75%, 77.2%, and 63.4%, respectively.

In [15], a method for the detection and recognition of paddy rice leaf diseases using SVM classifier was presented. Four diseases were detected and recognized with 98.3% accuracy. The proposed system takes images collected from surrounding agricultural lands as input and introduces image processing methods to extract features required for further processing. The k-means algorithm was used for image segmentation and SVM for disease recognition.

In [16], rice diseases were detected through leaves by applying SVM. The project was carried out in three stages: image acquisition through a digital camera, image segmentation, and leaf region segmentation using the k-means algorithm. Finally, SVM was used to classify leaf diseases with an accuracy of up to 90%.

In [17], a system was proposed for tomato disease classification using machine learning techniques and image-processing methods. Different algorithms such as random forests, SVM, decision tree, k-nearest neighbors, and naive Bayes were implemented, and the results were analyzed to find the best algorithm. The particular diseases considered are leaf curl, Septoria leaf spot, bacterial leaf spot and Alternaria. The dataset used consists of 1,090 images of infected tomato leaves in real-time acquired by phone camera at different infection stages, illumination (lighting), time, temperature, humidity levels, and location. After preprocessing the data, the author trained and tested models based on different algorithms. The results show that random forests achieve a better accuracy of 89% compared to other algorithms.

In their study, Kasinathan et al [18] proposed a model for pest detection using machine learning algorithms and 9-fold cross-validation to improve performance. The Wang dataset containing 225 images and nine classes of insects, and the Xie dataset containing 785 images and 24 classes of insects are used in this study. After data preprocessing, they trained and tested different models such as artificial neural network (ANN), SVM, k-nearest neighbors, naive Bayes and convolutional neural network (CNN). The highest classification rates of 91.5 % with Wang data and 90% with Xie data were achieved using the CNN model.

In their study, Marković et al [19] proposed a method to predict the daily appearance of insects during a season by taking into account temperature and relative humidity and applying machine learning for data processing and results generation. The data used consists of data on Helicoverpa armigera insects caught in traps equipped with light lamps collected from 17 localities in northern Serbia, Vojvodina province from 2019 to 2020 and daily data on the number of trapped insects, temperature, and relative humidity over the season. They used several machine learning algorithms, namely: K-nearest neighbors, Support Vector Machines with kernel = 'rbf' (RBF SVM), Support Vector Machines with kernel = 'poly' (Poly SVM), decision tree, random forest, multilayer perceptron (Neural Net), Ada Boost, Gaussian naive Bayes (G naive Bayes) and quadratic discriminant analysis (QDA) to create several varied models. The models were evaluated based on accuracy and confusion matrix according to their ability to predict the appearance of insects over different periods (1 day, 2 days, 3 days, etc.). The results show that Ada Boost performs best, with a detection accuracy of 86.3% for five days and a false detection rate of 11%.

The study conducted in [20] consists of developing a prediction model capable of estimating the level of damage caused by a pest using machine learning techniques, namely logistic regression and support vector machine (SVM). The study focused specifically on the pest "Chaetanaphothrips sp" and the banana crop in the city of Buenos Aires, Morropon, Piura, Peru. The data used contains 68 random samples of 25 banana plants taken from the crop (twice a week), the number of insects counted, the climate data and the soil data collected between November 2019 and August 2020. The initial dataset contains 23 attributes and the Principal Component Method (PCA) is used to select the 6 best attributes, which are: maximum temperature (°C), relative humidity (%), wind speed (m/s), evapotranspiration (mm), atmospheric pressure (MB) and the incidence of previous measurements. The performance of the models was evaluated based on accuracy, precision, and recall. The results obtained show that SVM offers the best performance with an accuracy of 79%, a precision of 100%, and a recall of 73% compared to logistic regression which is 64%, 71%, and 63% respectively.

In [21], a study was proposed to predict the infestation of cotton crops by the pest "spodoptera littoralis" using machine learning techniques such as random forest, ExtraTree, XGBoost, logistic regression, and linear regression. The data used consists of 130 records collected per week between September 2017 and February 2020 in the Nabat commercial hydroponic greenhouse in Al Mansouryah, Giza, Egypt. In addition, among other characteristics, temperature and relative humidity were also collected over the entire study period. The dataset has features such as infestation severity, highest temperature, lowest temperature, relative humidity, biological control protocol, number of thrips and Indicator for the use of fertilizers. The models were evaluated based on RRMSE and MAPE. The results show that XGBoost offers the best performance with RRMSE = 25.61% and MAPE = 17.79% when all attributes are taken into account.

The approach proposed in [22] aims to predict the appearance of Cotton insects and diseases using LSTM. They first formulated the problem as a time series prediction before applying LSTM to solve it. They used data from the Crop Pest Decision Support System, which consists of information on 10 Cotton pests and diseases recorded per week, along with corresponding weather conditions at 6 major locations in India. They compared the model's performance with other methods such as KNN, SVM, and random forests based on accuracy, area under the curve (AUC), and F1-score. The results show that LSTM performs best, with an accuracy of 91.7%, an AUC of 96.9% and an F1-score of 86%.

3. Methodology

In this section, we present the experimental design used to compare different multilabel classification methods and implement the proposed model. First, the datasets used in this study and the preprocessing techniques applied to clean and adapt the data are described. Then, a brief overview is given on the multilabel classification and the different algorithms tested. Finally, the evaluation metrics applied to evaluate multilabel classification models are presented.

3.1. Data collection

Two data sets were used in this study. The first dataset used consists of real historical data on crop infestation (millet, sorghum, maize, cowpea and groundnut) by pests in Niger from 2006 to 2022. The data was collected from the Directorate General of Plant Protection (DGPV) in Niamey, Niger. The dataset consists of 11,328 rows and 7 columns (Table 1), namely: Year, Month, Region, Department, Locality, Crop and Pests. The dataset includes 21 different classes of pests that are harmful to crops, such as Dereodus, Thrips, Grasshoppers, Rodents, Bugs, Aphidis, Sed-eating brids, Mylabris, Iules, Flower insects, Stem borers, Pod borer, Beetles, Leafinoppers, Hairy caterpillars, Ear head caterpilars, Fall armyworms, Defoliating caterpillars, Collar caterpillars, Midge and Mites.

The second dataset used consists of monthly and annual climatology data for the last 41 years (from 1981 to 2022) specific to the different areas infested by pests, obtained from the NASA website (https://power.larc.nasa.gov) through monthly API calls and using the geographic coordinates (longitude and latitude) of the infested areas as parameters. The data consist of 130,032 rows and 14 columns (Year, Month, Region, Department, Locality, Longitude, Latitude, Average Temperature, MAX Temperature, MIN Temperature, Precipitation, Relative Humidity, Surface Pressure and Wind Speed).

3.2. Data preprocessing

Both datasets used in this study present some challenges such as non-numeric attribute values, missing values and unimportant

Table 1. Overview of crop pest data

Year	Month	Region	Department	Locality	Crop	Pests
2022	8	Agadez	Iferouane	Iferouane	Corn	Beetles
2022	8	Agadez	Tchirozerine	Agadez	Corn	Grasshoppers
2022	9	Agadez	Tchirozerine	Dabaga	Corn	Grasshoppers
2022	9	Agadez	Tchirozerine	Dabaga	Corn	Thrips
2022	9	Agadez	Aderbissinat	Aderbissinat	Cowpea	Aphids
2022	7	Diffa	Goudoumaria	Goudoumaria	Millet	Grasshoppers
2022	7	Diffa	Maine-Soroa	Maine Soroa	Cowpea	Hairy caterpillars
2022	8	Diffa	N'Guigmi	N'Guigmi	Millet	Grasshoppers
2022	8	Diffa	N'Guigmi	Kablewa	Millet	Grasshoppers
2022	8	Diffa	Diffa	Chetimari	Millet	Grasshoppers
2022	8	Diffa	Diffa	Chetimari	Cowpea	Hairy caterpillars
2022	9	Diffa	N'Guigmi	Kablewa	Millet	Grasshoppers
2022	7	Dosso	Gaya	Bengou	Millet	Flower insects

columns which are not suitable for the multilabel classification task. Also, the sample number of pest classes is unbalanced.

Data preprocessing is an essential task that involves cleaning data and adapting it to a suitable format that can be used by algorithms. To meet these challenges, preprocessing is necessary. To achieve this, the following actions are carried out:

i. The values of attributes such as Region, Department, and Locality are strings and, therefore, cannot be read by certain machine learning algorithms. To avoid problems, these attributes are encoded into numeric values using the LabelEncoder tool in python, which consists of converting each value in the column into a number;
ii. All missing values are removed;
iii. Non important columns are removed;
iv. This work involves predicting the pests likely to appear over a period (months or years) using multilabel classification. The data to be used in this kind of problem is not linear and need to be transformed and adapted. To do this, a new dataset (S1) is created with 21 additional columns corresponding to the 21 pests present in the data. Then, using a newly created function, the data are grouped by year, region, department, commune, and crop while marking the column corresponding to the pest as 1 if it appears during the year in the locality and on the crop, and 0 otherwise. The rainy season extends from June to October. In the climate data, the data corresponding to these months are extracted in order to create a new dataset (S2) consisting of the locality annual values of cumulative precipitation, average relative humidity, average temperature, maximum temperature, and minimum temperature. Finally, S1 and S2 datasets are combined to create the SF1 dataset, which is used to predict pests over a year. To predict pests over a month, the same procedure is used to create another new SF2 dataset by aggregating the data by month instead of by year.

3.3. Multilabel classification methods

The main difference between traditional classification (i.e., multi-class or binary) and multilabel classification is the expected output of the trained models. Unlike a traditional classifier, which will only return a single output value, a multilabel classifier must produce a vector or subset of output values. According to [2] and [23], multilabel classification problems use three (3) different approaches:

- The transformation approach: it is one of the first approaches to carry out multilabel classification. It transforms the learning task into one or more binary or multi-class classification tasks by creating a new dataset from the set of original data so that traditional algorithms can be used to solve the problem;
- The adaptation approach: it extends traditional algorithms to process multilabel data directly and provide several labels instead of just one.
- The ensemble approach: it aims to correct machine learning problems such as overfitting, underfitting, label imbalance, or numerical anomalies linked to the order of labels, from which the two previous approaches suffer.

In the present study, for performance comparison purposes, a total of 9 multilabel classifiers are implemented and tested. They are distributed as follows:

- For the transformation approach, four (4) methods are selected, namely: Binary Relevance (BR), Classifier Chain (CC), Label Powerset (LP), and Pruned Sets (PS);
- For the adaptation approach, two (2) methods are used: BRkNN and ML- kNN;
- Finally, for the ensemble approach, three (3) methods are used: Ensemble of Classifier Chains (ECC), Ensemble of Pruned Sets (EPS), and RAndom k labEL sets (RAkEL).

3.4. Evaluation metrics

The evaluation of multilabel classification algorithms requires different metrics than those used in traditional classification because the output of any multilabel classifier consists of a predicted set of labels for each test instance. Moreover, unlike in a traditional scenario where with a single output class the prediction can only be correct or incorrect, a multilabel prediction, on the other hand, can be entirely correct, partially correct/incorrect (to different degrees), or completely incorrect [2]. For this reason, several metrics have been proposed specifically to evaluate multilabel classifiers, the most commonly used are:

- Hamming Loss (HL): this is the most commonly used metric. It is defined as the fraction of labels that are incorrectly predicted [24]. The main advantages of Hamming Loss are its simplicity and the fact that it is less affected by class imbalance because it considers the error for each label independently.
- Accuracy: this is the proportion of correct predictions. That is, the proportion between the number of correctly predicted labels and the total number of active labels, both in the actual label set and the predicted one [2]. It is calculated as the ratio of true positives (TP) and true negatives (TN) to the total number of samples. However, it contains weaknesses for a multilabel prediction evaluation. It does not capture the notion that the predicted subset may be partially correct/incorrect. Moreover, in such a problem where the classes are unbalanced, the accuracy can be misleading if the model mostly predicts the frequent labels correctly but fails on rare ones. Consequently, it is not used in the present study;
- Precision (P): it is the proportion of true positive predictions among all positive predictions made by the model. It is calculated as the ratio of TP to the sum of TP and false positives (FP). This is the most intuitive metric [2];
- Recall (R): it is the proportion between the number of true positive predictions and all actual positive instances. It is calculated as the ratio of TP to the sum of TP and false negatives (FN).
- F- Measure (F1): it is the harmonic average of precision and recall [2];

4. Results

Table 2. Comparison of model performance on monthly predictions

Model	Recall	Precision	F- Measure	Hamming Loss
BR	59.96	75.55	66.85	3.65
CC	61.99	73.38	67.21	3.72
LP	65.5	69.17	67.28	3.91
PS	66.12	70.14	68.07	3.81
BRkNN	45.91	45.07	45.49	6.75
ML- kNN	45.91	45.07	45.49	6.75
ECC	61.93	55.81	58.71	5.35
EPS	72.71	61.53	66.65	4.47
RAkEL	62.23	74.49	67.81	3.63

Table 3. Comparison of model performance on annual predictions

Model	Recall	Precision	F- Measure	Hamming Loss
BR	55.78	70.58	62.31	5.18
CC	56.02	69.53	62.05	5.27
LP	59.68	64.45	61.97	5.63
PS	60.98	65.97	63.38	5.42
BRkNN	43.66	41.96	42.79	8.97
ML- kNN	43.66	41.96	42.79	8.97
ECC	65.05	51.42	57.43	7.41
EPS	66.92	61.1	63.88	5.81
RAkEL	56.75	71.01	63.09	5.1

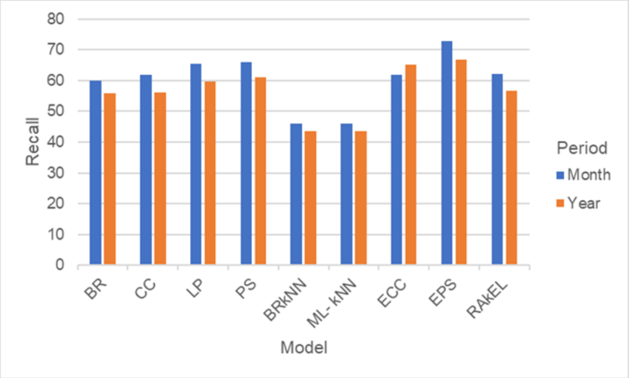

Fig. 1 Recall comparison of all models based on period.

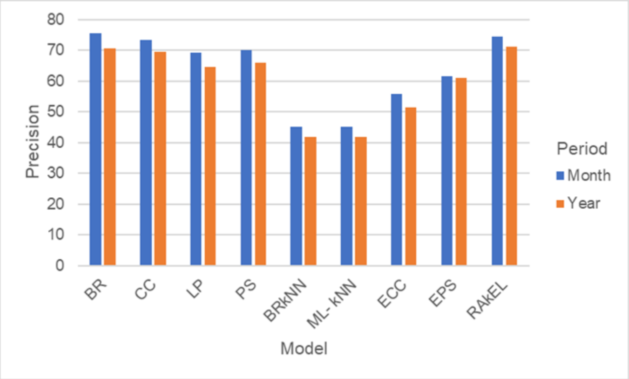

Fig. 2 Precision comparison of all models based on period.

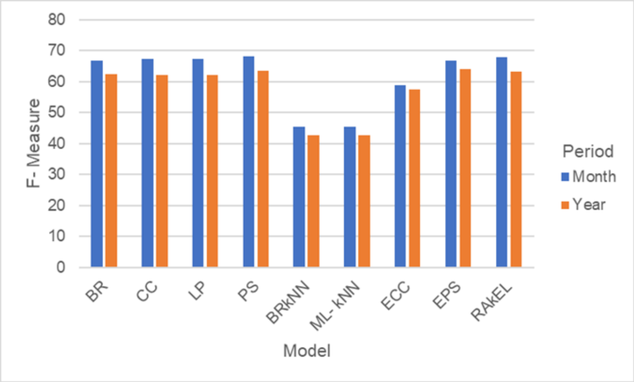

Fig. 3 F-Measure comparison of all models based on period.

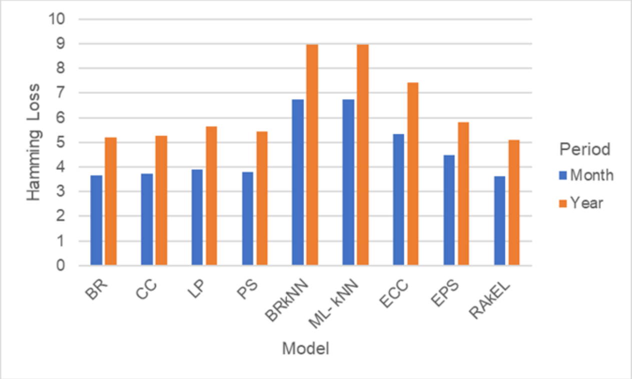

Fig. 4 Hamming Loss comparison of all models based on period.

5. Discussion

In this section, the experimental results are presented. For each model, the critical values of the evaluation metrics obtained from the test on the same data sets are discussed for comparison purposes.

Nine (9) multilabel classification models were implemented and evaluated over two different periods (monthly and annually) using SF2 and SF1 datasets, respectively. The two datasets are first divided into two subsets each, namely, a training set consisting of 80% of the data and a testing set consisting of the remaining 20% of the data. Then, the models are trained using K-fold cross-validation (with K = 5) to avoid overfitting. Since our dataset is imbalanced, we used the stratified K-fold method, as it is a technique that preserves the imbalanced class distribution in each fold by ensuring that each fold maintains the same class distribution as the original dataset [25]. This technique helps in mitigating the problems caused by imbalanced data and allows for more robust evaluation of models. In addition, all BR, CC, LP, PS, and RAkEL classifiers are trained and tested using the random forest algorithm as base classifier. The model performance evaluation metrics used include precision (P), recall (R), F1 score (F1), and Hamming Loss (HL). The evaluation results are shown in Table 2 and Table 3.

Considering the monthly forecasts, as shown in Table 2, we observe that:

i. In terms of recall, EPS outperforms all other models with a recall of 72.71%, followed by PS (66.12%) and LP (65. 5%). BRkNN and ML-KNN have the lowest recall (45.91%) while BR, ECC, CC, and RAkEL offer recalls between 59% and 63%. This means that, compared to other models, EPS maximizes the number of true positives. But this does not give any information about the prediction quality of the models on the true negatives. This information is given by the precision;
ii. The BR model offers the highest precision of up to 75.55%, followed by RAkEL (74.49%) and CC (73. 38%). However, recall or precision does not allow a model to be fully evaluated. Separately, these two measures are almost useless because, if the model predicts "positives" every time, the recall will be high, and if the model never predicts "positives", the precision will in turn be high. This indicates that the model is efficient while, on the contrary, it will be naiver than intelligent. The F1 score is introduced to make a good evaluation of model performance by combining recall and precision;
iii. PS offers the highest F1 score of 68.07%, followed by RAkEL, LP, CC, BR, EPS, and ECC with scores of 67.81%, 67.28%, 67.21%, 66.85%, 66.65%, and 58.71%, respectively;
iv. The BRkNN and ML-kNN models offer low F1 scores (45.49% each), so they are too bad;
v. For Hamming Loss analysis, RAkEL makes fewer poor label predictions (3.63%) compared to BR, CC, PS, LP, and EPS which are 3.65%, 3.72%, 3.81%, 3.91%, and 4.47% respectively. ECC, BRkNN, and ML-kNN provide high rates of label misprediction (above 5%);
vi. Although, PS offers the best F1 score compared to RAkEL, the latter makes fewer false label predictions. Both models can be used for monthly pest prediction.
vii. RAkEL was selected as the best model because of its fewer false label predictions.

Regarding annual forecasts (see Table 3), we observe:

i. A decrease in recall, precision, and F1 score, and an increase in Hamming Loss for all models;
ii. Globally, the prediction quality of the models decreases with an increase in the prediction scale;
iii. Based on the F1 score and Hamming Loss performance metrics, EPS has the highest F1 score (63.88%) followed by PS (63.38%), RAkEL (63.09%), BR (62.31%), CC (62.05). %), LP (61.97%), ECC (57.43%), BRkNN (42.79%) and ML-kNN (42.79%). Hower, RAkEL has the lowest Hamming Loss (5.1%) compared to BR (5.18%), CC (5.27%), PS (5.42%), LP (5.63%), EPS (5.81%), ECC (7.41%), BRkNN (8.97%) and ML-kNN (8.97%);
iv. RAkEL, EPS, and PS can equally be used for annual pest prediction;
v. RAkEL outperforms the other models because of its minimal Hamming Loss.

For instance, one of the main reasons why BRkNN and ML-kNN may not be a good predictor is that it treats each label as an independent binary classification problem. When a label is rare (i.e., a minority class like pest Iules), the k-nearest neighbors are more likely to belong to the majority class, making it challenging to predict the minority class accurately.

Fig. 1–4 show that the longer the prediction period, the poorer the models perform. This shows that the use of aggregated data leads to a loss of information, because climatic variations occurring during the year are not taken into account. These results suggest using monthly or aggregated data while retaining full information on climate variations occurring throughout the year to ensure better performance.

These results show that multilabel classification can be used to predict all crop pests over a period of one month or one year with great precision. Therefore, they contribute to the growing body of literature on the use of Machine Learning techniques (particularly multilabel classification) for crop pests' prediction to address food security and resource allocation challenges in Niger and elsewhere. However, in terms of innovation, this is one of the first study evaluating the effectiveness of several multilabel classification algorithms for predicting crop pests in Niger. Therefore, this study can serve as a basis for further research in this area, leading to more accurate and efficient methods for predicting crop pests in Niger and other countries.

Our study has some limitations. Firstly, the size of the dataset used is small. Although data augmentation can be an applicable solution, collecting more data is the ultimate solution. Secondly, the data are unbalanced, this can impact the performance of the model and penalize the prediction of minority classes. A practical solution is the use of others class balancing techniques.

6. Conclusion

In this study, we proposed an application of multilabel classification for the prediction of crop pests in Niger. We implemented and tested nine (9) multilabel classification methods based on the random forest algorithm as a base classifier. Our results show that RAkEL is a promising method which can be as a reference approach to implementing pest prediction systems. The proposed model helps in reducing pest harmful impact on crops and agricultural production.

In our future work, we will first focus on improving the model's performance on medium amounts of data. Then, we will introduce other multilabel classification methods. Secondly, we will try to integrate other data sources and parameters such as soil data to increase the number of parameters in order to make the model more robust and efficient. Finaly, we will discuss the implications in real-world applications. The model will be deployed at FUMA Gaskiya federation and tested first by more than 10,000 producers who are members of the FUMA Gaskiya federation in Maradi region - Niger before being opened to the general public. This will allow the model to be tested and validated in a real environment. It will be loaded via smartphones and without requiring an internet connection.

The main limitations in the adoption of this technology are the low level of education of Nigerian farmers and the lack of knowledge of the importance of this type of technology in their agricultural activities. But, this problem can be solved through communication and awareness raising among producers and the development of the application through a simple, user-friendly, easy-to-use interface available in different local languages.

.

Acknowledgements

This research is supported by Abdou Moumouni University, Niamey, Niger. Data was obtained from Directorate General of Plant Protection (DGPV) Niamey, Niger and NASA website.

Author contributions

Mahaman Lawali INOUSSA GARBA: Participated in all experiments and the design of the research plan, organized the study, participated in data collection and data processing, participated in data-analysis and contributed to the writing of the manuscript.
Harouna NAROUA: Coordinated and designed the research plan, participated in all the experiments, and contributed in the writing of the manuscript.
Chaibou KADRI: Participated in all the experiments and contributed in the writing of the manuscript.
Madougou GARBA: Participated in data collection and data processing, contributed to the writing of the manuscript.
Maman Aminou ALI: Contributed in the writing of the manuscript.

Conflicts of interest

None of the authors have conflict of interest related to the research and results presented in this paper.

References

[1] J. Bogatinovski, L. Todorovski, S. Džeroski, and D. Kocev, ‘Comprehensive comparative study of multi-label classification methods’, Expert Syst. Appl., vol. 203, p. 117215, Oct. 2022, doi: 10.1016/j.eswa.2022.117215.

[2] F. Herrera, F. Charte, A. J. Rivera, and M. J. del Jesus, ‘Multilabel Classification’, in Multilabel Classification : Problem Analysis, Metrics and Techniques, F. Herrera, F. Charte, A. J. Rivera, and M. J. del Jesus, Eds., Cham: Springer International Publishing, 2016, pp. 17–31. doi: 10.1007/978-3-319-41111-8_2.

[3] Z. A. Habou, M. K. Boubacar, and T. Adam, ‘Les systèmes de productions agricoles du Niger face au changement climatique: défis et perspectives’, Int. J. Biol. Chem. Sci., vol. 10, no. 3, pp. 1262–1272, 2016.

[4] H. Kaur, D. D. Prashar, and Madhuri, ‘Applications of Machine Learning In Plant Disease Detection’, Think India J., vol. 22, no. 17, Art. no. 17, Sep. 2019.

[5] I. Bhakta, S. Phadikar, K. Majumder, H. Mukherjee, and A. Sau, ‘A novel plant disease prediction model based on thermal images using modified deep convolutional neural network’, Precis. Agric., vol. 24, no. 1, pp. 23–39, Feb. 2023, doi: 10.1007/s11119-022-09927-x.

[6] G. Delnevo, R. Girau, C. Ceccarini, and C. Prandi, ‘A Deep Learning and Social IoT Approach for Plants Disease Prediction Toward a Sustainable Agriculture’, IEEE Internet Things J., vol. 9, no. 10, pp. 7243–7250, May 2022, doi: 10.1109/JIOT.2021.3097379.

[7] S. Lee and C. M. Yun, ‘A deep learning model for predicting risks of crop pests and diseases from sequential environmental data’, Plant Methods, vol. 19, no. 1, p. 145, Dec. 2023, doi: 10.1186/s13007-023-01122-x.

[8] Y. Xu et al., ‘A deep learning model for rapid classification of tea coal disease’, Plant Methods, vol. 19, no. 1, p. 98, Sep. 2023, doi: 10.1186/s13007-023-01074-2.

[9] E. Ayan, ‘Genetic Algorithm-Based Hyperparameter Optimization for Convolutional Neural Networks in the Classification of Crop Pests’, Arab. J. Sci. Eng., vol. 49, no. 3, pp. 3079–3093, Mar. 2024, doi: 10.1007/s13369-023-07916-4.

[10] S. Ramesh et al., ‘Plant Disease Detection Using Machine Learning’, in 2018 International Conference on Design Innovations for 3Cs Compute Communicate Control (ICDI3C), Apr. 2018, pp. 41–45. doi: 10.1109/ICDI3C.2018.00017.

[11] K. Ahmed, T. Shahidi, S. Irfanul Alam, and S. Momen, ‘Rice Leaf Disease Detection Using Machine Learning Techniques’, Dec. 2019, pp. 1–5. doi: 10.1109/STI47673.2019.9068096.

[12] S. Mishra, R. Sachan, and D. Rajpal, ‘Deep Convolutional Neural Network based Detection System for Real-time Corn Plant Disease Recognition’, Procedia Comput. Sci., vol. 167, pp. 2003–2010, Jan. 2020, doi: 10.1016/j.procs.2020.03.236.

[13] S. Gnanasekaran and G. Opiyo, ‘A predictive machine learning application in agriculture: Cassava disease detection and classification with imbalanced dataset using convolutional neural networks’, Egypt. Inform. J., vol. 22, Mar. 2020, doi: 10.1016/j.eij.2020.02.007.

[14] M. Agarwal, A. Singh, S. Arjaria, A. Sinha, and S. Gupta, ‘ToLeD: Tomato Leaf Disease Detection using Convolution Neural Network’, Procedia Comput. Sci., vol. 167, pp. 293–301, Jan. 2020, doi: 10.1016/j.procs.2020.03.225.

[15] T. Devi, P. Neelamegam, and A. Srinivasan, ‘Plant Leaf Disease Detection using K means Segmentation’, 2018. Accessed: May 28, 2024. [Online]. Available: https://www.semanticscholar.org/paper/Plant-Leaf-Disease-Detection-using-K-means-Devi-Neelamegam/c361350782367f0473f89a069ff4b269dc218ba1

[16] M. Shanthalakshmi, M. Sandhiya, M. Rajalakshmi, and V. Ratheesh, ‘Paddy Disease Detection and Pesticide Recommender System for Farmers Using Multi SVM Technique’, Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol., pp. 721–725, Mar. 2019, doi: 10.32628/CSEIT1952214.

[17] Neelakantan . P, ‘Analyzing the best machine learning algorithm for plant disease classification’, Mater. Today Proc., vol. 80, pp. 3668–3671, Jan. 2023, doi: 10.1016/j.matpr.2021.07.358.

[18] T. Kasinathan, D. Singaraju, and S. R. Uyyala, ‘Insect classification and detection in field crops using modern machine learning techniques’, Inf. Process. Agric., vol. 8, no. 3, pp. 446–457, Sep. 2021, doi: 10.1016/j.inpa.2020.09.006.

[19] D. Marković, D. Vujičić, S. Tanasković, B. Đorđević, S. Ranđić, and Z. Stamenković, ‘Prediction of Pest Insect Appearance Using Sensors and Machine Learning’, Sensors, vol. 21, no. 14, p. 4846, Jul. 2021, doi: 10.3390/s21144846.

[20] E. Almeyda, J. Paiva Mimbela, and W. Ipanaqué, ‘Pest Incidence Prediction in Organic Banana Crops with Machine Learning Techniques’, Oct. 2020, pp. 1–4. doi: 10.1109/EIRCON51178.2020.9254034.

[21] A. Tageldin, D. Adly, H. Mostafa, and H. S. Mohammed, ‘Applying Machine Learning Technology in the Prediction of Crop Infestation with Cotton Leafworm in Greenhouse’. bioRxiv, p. 2020.09.17.301168, Sep. 19, 2020. doi: 10.1101/2020.09.17.301168.

[22] Q. Xiao, W. Li, P. Chen (陈鹏), and B. Wang, ‘Prediction of Crop Pests and Diseases in Cotton by Long Short Term Memory Network’, 2018, pp. 11–16. doi: 10.1007/978-3-319-95933-7_2.

[23] P. Szymański and T. Kajdanowicz, ‘scikit-multilearn: A Python library for Multi-Label Classification’, J. Mach. Learn. Res., vol. 20, no. 6, pp. 1–22, 2019.

[24] Y. Ren et al., ‘Multi-label classification for multi-drug resistance prediction of Escherichia coli’, Comput. Struct. Biotechnol. J., vol. 20, pp. 1264–1270, 2022, doi: 10.1016/j.csbj.2022.03.007.

[25] M. T r, V. K. V, D. K. V, O. Geman, M. Margala, and M. Guduri, ‘The stratified K-folds cross-validation and class-balancing methods with high-performance ensemble classifiers for breast cancer classification’, Healthc. Anal., vol. 4, p. 100247, Dec. 2023, doi: 10.1016/j.health.2023.100247.