Available online to ejournal.pnc.ac.id

Journal of Innovation Information Technology and Application

Journal Page is available to https://ejournal.pnc.ac.id/index.php/jinita

From Text to Insights: NLP-Driven Classification of Infectious Diseases Based on Ecological Risk Factors

Saviour Inyang [1], Imeh Umoren [2]

[1,2] Department of Computer Science, Akwa Ibom State University, Mkpat Enin, Nigeria

email:[1] sinyang5672@gmail.com, [2] imehumoren@aksu.edu.ng

ARTICLE INFO

Article history:
Received 25 September 2023
Revised 28 November 2023
Accepted 21 December 2023
Available online 30 December 2023

Keywords:
Classification,
Natural Language Processing,
Ecology,
Infectious Disease

IEEE style in citing this article:
S. Inyang and I. Umoren , "From Text to Insights: NLP-Driven Classification of Infectious Diseases Based on Ecological Risk Factors ," Journal of Innovation Information Technology and Application (JINITA), vol. 5, no. 2, pp. 154–165, Dec. 2023.

ABSTRACT

Numerous factors can affect the development of infectious diseases that emerge. While many are the result of natural procedures, such as the gradual emergence of viruses over time, certain ones are the result of human activity. Human activities form an integral part of our ecosystem, and especially the ecological aspect of human activities can encourage disease transmission. Additionally, Health ecologists examine changes in the biological, physical, social, and economic settings to understand how these alterations impact the mental and physical well-being of individuals. Hence, this research adopts a Framework-Based Method (FBM) in carrying out the task of classification of infectious diseases. The Framework-Based Method outlines all phases that this research follows to carry out the infectious disease classification process, providing a structured and reproducible approach. Results show that: XGB: Confusion matrix accuracy: 76%, Kappa: 73%, RF: Confusion matrix accuracy: 65%, Kappa: 60%, SVM: Confusion matrix accuracy: 63%, Kappa: 58%, ANN: Confusion matrix accuracy: 71%, Kappa: 67%, LDA: Confusion matrix accuracy: 76%, Kappa: 73%, GBM: Confusion matrix accuracy: 60%, Kappa: 53%, KNN: Confusion matrix accuracy: 43%, Kappa: 34%, and DT: Confusion matrix accuracy: 37%, Kappa: 29%. Furthermore, a Deep Learning model BERT was integrated with the best classification model XGBoots to create an interactive interface for users to carry out infectious disease classification. This integration enhances user experience and accessibility, contributing to the practical application of machine learning and Natural language processing in ecological disease classification.

1. INTRODUCTION

Numerous factors can affect the development of infectious diseases that emerge. While many are the result of natural procedures, such as the gradual emergence of viruses over time, certain ones are the result of human activity [1]. Several reasons have contributed to these changes, including population growth, urbanization of rural areas, international air travel, global poverty, armed conflicts, and unfavorable changes in the environment brought on by economic growth and land use. Furthermore, Public health concerns are growing and getting more complex, in part because of the social and environmental risks posed by worldwide environmental alterations brought on by rapid industrialization, population growth, excessive exploitation of natural resources, and improper technological application [4]. The life-sustaining resources of the environment are being used unsustainably and in significant amounts. According to the Millennium Ecosystem Assessment, over the next 50 years, these disruptions might get worse and currently affect the well-being of individuals [2]. Human health and illness are impacted by numerous complex factors. Infectious diseases transmitted by humans and animals through interaction, contaminated environments, contaminated food, and contaminated water pose risks to public health. Since health is dynamic, varies over

**) Corresponding Author : sinyang5672@gmail.com*

time, and has many diverse elements, ecological opinions on food and the environment which include aquaculture, agriculture, and the total food systems are under a lot of strain [3].

In 2013, infectious diseases caused over 9 million fatalities and over 45 million years of lost productivity due to disability [5]. Nevertheless, Healthcare systems including the surrounding area are home to a wide variety of illnesses and microorganisms. Although viruses and bacteria vary greatly, the method by which germs spread from one individual to someone else is continuous within an environment. Therefore, exposure to environmental pollutants has been linked to a variety of human diseases and conditions such as infectious diseases. Finding disorders that may be related to environmental contaminants and determining the data sources that are already available regarding these diseases are essential steps in the effort to more precisely characterize links connecting environmental exposures and adverse health consequences [6].

On this note, this study investigates the diverse factors influencing the development of infectious diseases, distinguishing between natural and human-induced processes. Examining the ecological aspect of human activities to understand its role in encouraging disease transmission within the ecosystem. Also, develop and apply a Framework-Based Method (FBM) for the structured and reproducible classification of infectious diseases, encompassing data collection, preprocessing, and model training. Conduct a comparative analysis of classification models, evaluating their performance and the integration of Deep Learning model BERT with the best-performing classification model to create an interactive interface, enhancing user experience and accessibility in infectious disease classification.

2. RELATED LITERATURE

Many studies have been carried out on infectious diseases. We delve into the key findings and broader implications of these studies within the context of infectious disease research and its impact on public health. [7] examined Japan's infectious disease surveillance system, unveiling the significant role of legal amendments in reducing illness rates. However, their research's timeframe-specific nature limits the applicability of their findings. Nevertheless, it underscores the significance of policy adjustments in disease control. Also [8] embarked on a study to uncover the global trends in emerging infectious diseases (EIDs), underscoring the influence of socioeconomic, environmental, and ecological factors in the emergence of EIDs. Their work accentuates the necessity of a multidisciplinary approach to comprehend and mitigate EID risks. [9] proposed a theoretical framework for amalgamating data and models in infectious disease research. There was an emphasis on data gathering to enhance disease modeling while underscoring the critical importance of comprehensive data for precise predictions. Also, an assessment by [10] regarding China's capability to manage infectious diseases highlights the need for comprehensive prevention and response strategies. This study underscores the urgency of proactive measures to tackle future disease threats. It was [11] that investigated the utility of mathematical models in understanding the intricate dynamics of infectious diseases on a global scale. Their work elucidates the interconnectedness of infectious diseases and the potential for regional and global repercussions if interventions fall short. Again, the suggestion of leveraging mobile phone data to connect movement patterns to infectious diseases was presented in [12] introducing novel possibilities for characterizing population behavior and predicting disease outbreaks. This innovative approach may revolutionize disease tracking and response strategies. Furthermore, an exploration of the interplay between climate and infectious diseases, suggesting the potential for interdisciplinary cooperation between biology and climate research to gain deeper insights into disease dynamics was presented in [13]. Identifying dynamics can help create patterns in infectious disease occurrence, therefore a study conducted by [14] centers on identifying patterns in the occurrence of infectious disease syndromes in Mongolia. Their application of predictive models to detect rising disease rates underscores the potential of syndrome-based assessments in forecasting disease trends. Furthermore, in the modeling of infectious disease, a study conducted by [15] offers an overview of modeling infectious disease transmission. Their discussion on incorporating intricate data and advanced inference techniques underscores the significance of adapting modeling approaches to real-world disease spread. Therefore, with the evolving landscape of global changes and their possible ramifications for infectious diseases. There is a call for research adaptation to underscore the need to anticipate and address emerging challenges in disease prevention and control [16]. Also [17] provides the basis for the classification of infectious diseases based on semantic natural language processing. This study forms the basis of this research work. This study was limited to using one machine algorithm in the semantic classification. Hence this research will expand

further by introducing more algorithms for the comparative classification of infectious diseases using Natural Language Processing through ecological risk factors.

3. METHOD

In this research, we used a Framework-Based Method (FBM) to carry out the task of classification of infectious diseases. A FBM is a widely used strategy in computer science research, where a structured system of concepts is employed to guide and enhance research studies [18]. In Figure 1 below the frameworks used in this research are presented. The framework in Figure 1 presents all the different components or constitutes that show the different processes that this research carried out from the point of data collection, creating of infectious disease corpus, preprocessing of the data, creation of the document term matrix, text analysis, and visualization, training, and testing, comparative analysis, performance evaluation and deployment of the model. Nevertheless, all these individual sections are available the framework is discussed below in this research.

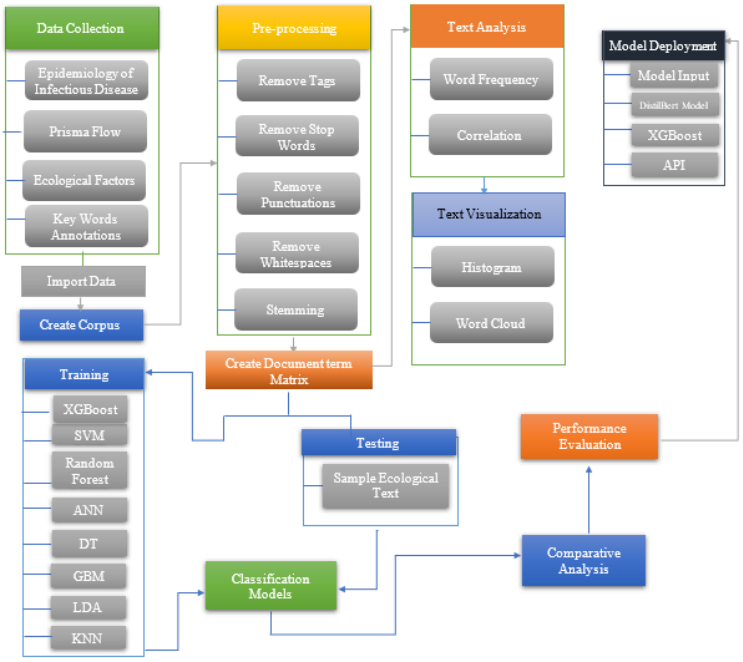

Figure 1. Framework for NLP-Driven Classification of Infectious Diseases

3.1 Data collection

Research is the process of gathering experiences or observations. Based on the data acquired, a researcher may assess their hypothesis. The three processes that were used to split data collection into this part are;

3.1.1 Epidemiology: Epidemiology examines the distribution (who, when, and where) and trends of health and illness situations in a certain population. As a data source for this study, we will use the epidemiological background of each disease that is transmissible.

3.1.2 Prisma Flow: A Prisma flow diagram was used in the representation of the literature that was selected from the epidemiology of infectious disease which is presented in [17]

3.1.3 Ecological Factors: The biotic and abiotic variables are represented by this. Anything that affects the natural world is considered to be an environmental or ecological factor. Environmental factors include things like water, air, soil, climate, local vegetation, and landforms. Environmental damage, forest loss, sewage contamination, warming temperatures, and climate change are the top five ecological problems affecting the well-being and health of humans. Therefore, we will identify all the ecological components that will make up a causative in the infectious disease given a given extracted epidemiology of the chosen infectious diseases from journal articles.

3.1.4 Keyword Annotation: In this study, we will extract and annotate all the keywords that represent ecological aspects for each infectious disease that is chosen. We will accomplish this by using headers and terms in bold and searching for the most significant points, arguments, and supporting evidence.

3.2 Corpus

Corpus construction entails creating a machine-readable text compilation that mirrors a specific language
or field. This process serves as a valuable resource for advancing and assessing natural language processing (NLP) algorithms and applications. In this research from the epidemiology data, we form bags of words that will be used in the NLP task

3.3 Data Preprocessing

Preprocessing data represents a data strategy that is commonly used to turn data into a beneficial and effective form. In this research, our processing phase follows; Removing Tags, Removing Stop Words, Removing Punctuation, removing whitespaces, and stemming which forms a standard preprocessing phase in NLP tasks.

3.4 Document Term Matrix

A document-term matrix or term-document matrix is a mathematical matrix that describes the frequency of terms that appear throughout a collection of documents. The documents in a document-term matrix are denoted by rows, and the terms are represented by columns. The document-phrase matrix is essentially a matrix that lists the frequency of each term over the whole corpus of written texts. In this research, we transform the corpus known as bags of words to represent each term in the matrix as a unique binary equivalent.

3.5 Text Analysis

The technique of analyzing and understanding human-written text using computer technologies to generate business insights is known as text analysis. Text analysis software can automatically classify, sort, and extract data from texts to discover patterns, connections, emotions, and other important information. We will employ word frequency in this study, which involves measuring the number of times a specific word is found in a text or collection of texts. Term frequency evaluates the relevance of each word.

3.6 Text Visualization

Text visualization is a means of visually presenting textual data using graphs, charts, or word clouds. This summarizes the content, detects trends and patterns across documents, and provides quick access to the most significant terms in a text. Word clouds are a fantastic place for beginning when displaying qualitative data.

They can be used for exploratory research to identify what can be found in a data set to develop labeling requirements for more extensive text analysis and visualization, as well as to provide basic quick insights.

3.7 Training and Testing

The main difference between training data and testing data is that the previous type is a subset of the original data used to train the machine learning model, whereas the second is used to evaluate the model's correctness. The training dataset is often larger than the testing dataset. Train and test datasets are often divided 80:20, 70:30, or 90:10. In this research we trained all the eight machine learning models in Figure 1 based on the document term matrix created earlier in D.

3.8 Classification

The most common Machine Learning technique is categorization, which builds a model from a set of pre-classified instances that can classify the whole set of records. This technology is especially well-suited for applications such as medical records and sickness risk analysis. In this research, we employed eight machine learning algorithms which are: XGBoost, Support vector machine, Random Forest, Artificial neural network, Decision Tree, Gradient Boosting Machine, Linear Discriminant Analysis, and K-nearest Neighbor were all training data which the document term matrix

3.9 Performance Evaluation

In machine learning, the effectiveness of machine learning models is determined using performance assessment measures or metrics. This helps us determine how well our machine-learning model will perform on a dataset it has not seen before. When analyzing the performance of machine learning models on new datasets, performance evaluation criteria are critical. There's a good possibility the model will continually perform better on the dataset you trained it on. However, in this research, machine learning evaluation metrics such as confusion metrics and kappa statistics were adopted in the evaluation of the trained model's performance.

3.10Model Deployment

In this section, we deployed an interface using the DistilBert pre-trained model alongside our trained model Where a user can easily supply ecological sample text and classify infectious diseases.

4. RESULTS AND DISCUSSION

4.1 Exploratory Data Analysis

Exploratory Data Analysis (EDA) holds utmost significance in the realm of machine learning research as it involves an in-depth exploration and comprehension of a dataset before deploying any modeling techniques[21]. It facilitates researchers in acquiring profound insights, recognizing patterns, and detecting anomalies within the dataset. From the data gathered in this research [17], EDA was conducted on the data to get more insight into the data. Furthermore, from a total of 342 epidemiological articles that report ecological factors that affect infectious disease on the selected 9 diseases which are (Malaria, Tuberculosis, Measles, Polio, Avian_Influenza, and Cholera) from 1998 to 2022, figure 2: represents the ecological factors count based on the 9-disease selected in this research.

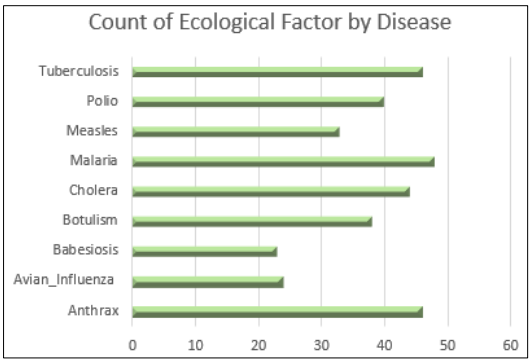

Figure 2. Ecological factors count based on disease.

Furthermore, some articles reported ecological factors based on their respective years of publications and it was observed that deforestation was mostly reported across the selected articles in the research from 1998 to 2022. Again, the total number of journals that reported ecological factors that affect infectious disease yearly, shows that over 112 journals from 2016 to 2017 recorded the highest ecological factors for infectious disease in the 9 selected diseases in this research.

4.2 Text Analysis and Visualizations

Text analysis and visualization are vital components of natural language processing (NLP) that play a crucial role in extracting valuable insights and understanding from unstructured text data. They provide essential techniques for processing and interpreting large amounts of textual information, enabling businesses and researchers to uncover patterns, trends, sentiments, and connections within the data. We present below the results of the text analysis using word frequency, world clouds, histograms, etc. Figure 3 represents word frequencies from the ecological sentences using a document term matrix for the text analysis.

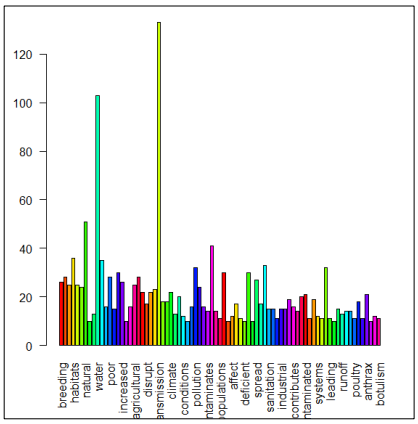

Figure 3. Word frequency based on ecological sentence

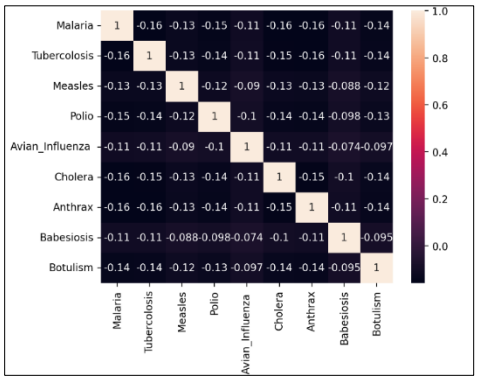

Figure 4. CM of infectious disease

Correlation analysis on the text-based classes of infectious disease in other to determine the level of correlation among the various classes of the disease using the correlation matrix (CM) which is depicted in Figure 4. Furthermore, text visualization was carried out using word clouds to detect the most frequent ecological words that were present in the sentences. Figure 5 represents ecological factors from sentences that describe infectious diseases.

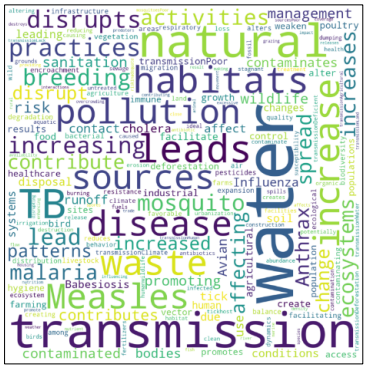

Figure 5. word cloud for ecological factors in infectious disease.

4.3 Classification

Machine learning, a subset of computer science, delves into the realm of algorithms that acquire knowledge from data examples. Classification, a key process within this field, involves applying machine learning techniques to decipher how and where to assign classifiers to instances within a given problem domain [19]. To illustrate, consider the straightforward task of classifying traffic as either "Yes" or "No" for a specific route. Within machine learning, a myriad of categorization tasks exist, each demanding distinct modeling approaches. However, the choice of the most suitable approach for a given classification task lacks a universal model-to-problem mapping. Instead, practitioners are encouraged to conduct controlled experiments to ascertain which algorithm and its configuration yield optimal results. The performance of classification predictive modeling algorithms hinges on the examination of their outcomes. Commonly, classification accuracy serves as a prevalent metric for assessing model performance, primarily relying on projected class labels. Hence, we present the classification process in this section below.

4.3.1 Model Training

Model training in machine learning is the procedure of instructing a mathematical model using a given dataset, enabling it to make precise predictions or decisions[20]. This training process involves exposing the model to labeled examples, enabling it to grasp the underlying patterns and relationships within the data. Throughout training, the model adjusts its internal parameters to minimize the disparity between its predictions and the actual labels. A total of 342 observations with 2 variables and 80% was used as the train set with a total of 273.6. Again, before we trained our model, to be able to classify

ecological text, we needed to preprocess our text data while removing stopwords and also looking at the term the sparsity of for the model where sparsity is good to yield statistical benefit for the model and also help and make it easily interpreted by human. Furthermore, we present the cross-section of the document term matrix that we use for the transformation of the ecological sentences that were present in the corpus. Figure 6 represents the document term matrix.

	breeding	deforestation	disrupts	habitats	increasing	mosquito	natural	mosquitoes	water	leads	management	poor	promoting	waste	increased	urbanization
1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
2	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
3	1	0	0	0	0	1	0	0	1	1	1	1	1	1	0	0
4	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1
5	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0
7	1	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0
8	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
9	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
10	1	1	0	0	0	1	0	0	0	0	0	0	0	0	1	0
11	1	0	0	0	1	1	0	0	1	0	0	0	0	0	0	0

Showing 1 to 11 of 342 entries, 50 total columns

Figure 6. Document Matrix.

4.3.2 Machine Learning Trained Infectious Disease Classifiers (XGBoots, Random Forest, Support Vector Machine , Artificial Neural Network, Decision Tree, Gradient boosting Algorithm, Linear Discriminant Analysis, K-Nearest Neighbor)

In this research, we carried out a comparative analysis of eight machine learning classifiers together were used in the training of document term matrices from NLP data. we present the comparative performance below;
From the use of the Xgboots Algorithm in training the data, we present the result in Figure 7.

```
eXtreme Gradient Boosting

273 samples
 71 predictor
  9 classes: 'Anthrax', 'Avian_Influenza ', 'Babesiosis', 'Botulism', 'Cholera', 'Malaria', 'Measles', 'Polio', 'Tubercolosis'

No pre-processing
Resampling: Cross-Validated (7 fold)
Summary of sample sizes: 233, 235, 233, 235, 235, 233, ...
Resampling results across tuning parameters:

  nrounds  Accuracy   Kappa
  100      0.7071718  0.6669999
  200      0.7030364  0.6622947
  300      0.6958936  0.6542596

Tuning parameter 'max_depth' was held constant at a value of 6
Tuning parameter 'eta' was held constant at a value
 held constant at a value of 1
Tuning parameter 'min_child_weight' was held constant at a value of 1
Tuning
 parameter 'subsample' was held constant at a value of 1
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were nrounds = 100, max_depth = 6, eta = 0.3, gamma = 0, colsample_bytree =
 1, min_child_weight = 1 and subsample = 1.
```

Figure 7. XGboost Training Results

Again, from the use of the Random Forest Algorithm in training the data, we present the result of the training in Figure 8.

```
Random Forest

273 samples
 71 predictor
  9 classes: 'Anthrax', 'Avian_Influenza ', 'Babesiosis', 'Botulism', 'Cholera', 'Malaria', 'Measles', 'Polio', 'Tubercolosis'

No pre-processing
Resampling: Cross-Validated (7 fold)
Summary of sample sizes: 231, 233, 233, 235, 236, 234, ...
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
   2    0.6266344  0.5707110
  36    0.7763971  0.7453723
  71    0.7550291  0.7212858

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 36.
```

Figure 8. Random Forest Training Results

.
Support Vector Machine Algorithm in training the data, we present the result of the training in Figure 9.

```
Support Vector Machines with Radial Basis Function Kernel

273 samples
 71 predictor
  9 classes: 'Anthrax', 'Avian_Influenza ', 'Babesiosis', 'Botulism', 'Cholera', 'Malaria', 'Measles', 'Polio', 'Tubercolosis'

No pre-processing
Resampling: Cross-Validated (7 fold)
Summary of sample sizes: 233, 234, 233, 236, 234, 234, ...
Resampling results across tuning parameters:

  C     Accuracy   Kappa
  0.25  0.3988046  0.3079432
  0.50  0.6623998  0.6137593
  1.00  0.6846674  0.6409620

Tuning parameter 'sigma' was held constant at a value of 0.007758208
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were sigma = 0.007758208 and C = 1.
```

Figure 9. SVM Training Results

We present the the neural network architecture which was trained on the ecological sample text in Figure 10.

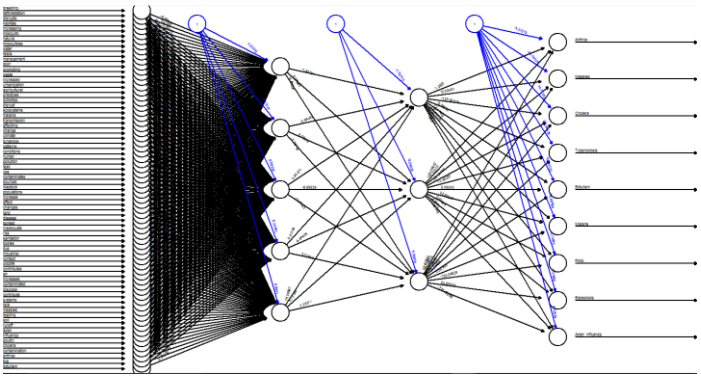

Figure 10. Architecture of the train neural network

Furthermore, the overall statistics accuracy of the confusion matrix is presented below in Figure 11.

```
Overall Statistics

               Accuracy : 0.7101
                 95% CI : (0.5884, 0.8131)
    No Information Rate : 0.1449
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6722

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Anthrax Class: Avian_Influenza  Class: Babesiosis Class: Botulism Class: Cholera Class: Malaria
Sensitivity              0.71429              0.8750           0.40000         0.33333         0.75000         0.71429
Specificity              0.90323              0.9836           0.98438         0.95238         0.98361         0.96774
Pos Pred Value           0.45455              0.8750           0.66667         0.40000         0.85714         0.71429
Neg Pred Value           0.96552              0.9836           0.95455         0.93750         0.96774         0.96774
Prevalence               0.10145              0.1159           0.07246         0.08696         0.11594         0.10145
Detection Rate           0.07246              0.1014           0.02899         0.02899         0.08696         0.07246
Detection Prevalence     0.15942              0.1159           0.04348         0.07246         0.10145         0.10145
Balanced Accuracy        0.80876              0.9293           0.69219         0.64286         0.86680         0.84101
                     Class: Measles Class: Polio Class: Tubercolosis
Sensitivity                  1.0000       0.55556               0.8000
Specificity                  0.9833       0.91667               1.0000
Pos Pred Value               0.9000       0.50000               1.0000
Neg Pred Value               1.0000       0.93220               0.9672
Prevalence                   0.1304       0.13043               0.1449
Detection Rate               0.1304       0.07246               0.1159
Detection Prevalence         0.1449       0.14493               0.1159
Balanced Accuracy            0.9917       0.73611               0.9000
```

Figure 11. Neural Network Accuracy

Furthermore, this research also carried out a comparative analysis using a decision tree classifier to give a comparison of the performance of each algorithm in the ecological classification of infectious diseases, we present the performance of the decision tree in Figure 12

```
CART

273 samples
 71 predictor
  9 classes: 'Anthrax', 'Avian_Influenza ', 'Babesiosis', 'Botulism', 'Cholera', 'Malaria', 'Measles', 'Polio', 'Tubercolosis'

No pre-processing
Resampling: Cross-Validated (7 fold)
Summary of sample sizes: 232, 234, 233, 235, 235, 234, ...
Resampling results across tuning parameters:

  cp          Accuracy   Kappa
  0.05128205  0.3995558  0.31604269
  0.08689459  0.3291875  0.23454749
  0.11111111  0.1644812  0.02706605

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was cp = 0.05128205.
```

Figure 12. Decision tree Accuracy

Again, this research also carried out a comparative analysis using the Gradient Boosting Machine classifier to give a comparison of the performance of each algorithm in the ecological classification of infectious diseases, we present the performance of the Gradient Boosting Machine in Figure 13.

```
Stochastic Gradient Boosting

273 samples
 71 predictor
  9 classes: 'Anthrax', 'Avian_Influenza ', 'Babesiosis', 'Botulism', 'Cholera', 'Malaria', 'Measles', 'Polio', 'Tubercolosis'

No pre-processing
Resampling: Cross-Validated (7 fold)
Summary of sample sizes: 232, 234, 233, 237, 235, 234, ...
Resampling results across tuning parameters:

  interaction.depth  n.trees  Accuracy   Kappa
  1                   50      0.5044143  0.4334889
  1                  100      0.5408488  0.4757608
  1                  150      0.5811958  0.5226203
  2                   50      0.5493050  0.4843170
  2                  100      0.5561913  0.4939403
  2                  150      0.5966511  0.5392883
  3                   50      0.5370939  0.4700033
  3                  100      0.5743221  0.5148496
  3                  150      0.5960917  0.5393070

Tuning parameter 'shrinkage' was held constant at a value of 0.1
Tuning parameter 'n.minobsinnode' was held constant at
 a value of 10
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were n.trees = 150, interaction.depth = 2, shrinkage = 0.1 and n.minobsinnode = 10.
```

Figure 13. Gradient Boosting Machine Accuracy

Again, this research also carried out a comparative analysis using a Linear Discriminant Analysis Machine classifier to give a comparison of the performance of each algorithm in the ecological classification of infectious diseases, we present the performance of Linear Discriminant Analysis in Figure 14.

```
Linear Discriminant Analysis

273 samples
 71 predictor
  9 classes: 'Anthrax', 'Avian_Influenza ', 'Babesiosis', 'Botulism', 'Cholera', 'Malaria', 'Measles', 'Polio', 'Tubercolosis'

No pre-processing
Resampling: Cross-Validated (7 fold)
Summary of sample sizes: 236, 234, 233, 234, 232, 234, ...
Resampling results:

  Accuracy   Kappa
  0.7696395  0.7381647
```

Figure 14. Linear Discriminant Analysis Accuracy

Nevertheless, this research also carried out a comparative analysis using the KNN Machine classifier to give a comparison of the performance of each algorithm in the ecological classification of infectious diseases, we present the performance of KNN in Figure 15.

```
No pre-processing
Resampling: Cross-Validated (7 fold)
Summary of sample sizes: 234, 233, 234, 235, 234, 234, ...
Resampling results across tuning parameters:

  k  Accuracy   Kappa
  5  0.4469684  0.3709591
  7  0.4760893  0.4027381
  9  0.4469636  0.3694215
```

Figure 15. KNN Accuracy

4.4 Comparative Analysis of Our Model

Additionally, a comparative tabulated analysis was carried out to compare different Classification algorithms used in this study, to observe the different accuracy variations based on the accuracy of the classification which is presented in Table 1. From Table 1 we can see that after the cross-validation accuracy, XGBoost has the highest percentage accuracy as compared to all other machine classification models.

Table 1: Comparative Analysis of Our Model

Model Type	Optimal number of rounds	Turning parameter & Value	Fold	Confusion Matrix	Kappa	Cross validated Accuracy
XGBoots	100 nrounds	Max.depth=6, eta=1,minchildweight=1, subsample=1	7	0.71	0.67	0.72
Random Forest	Mtry=36	Mtry=36	7	0.77	0.74	0.71
Support Vector Machine	Cost=1	Sigma= 0.00775	7	0.68	0.64	0.652
Artificial Neural Network	100 iterations		7	0.71	0.67	0.64
Decision Tree	Complexity Parameter =0.5128205	Complexity Parameter =0.5128205	7	0.37	0.29	0.4
Gradient boosting Algorithm	n.trees=150	Shinkage=0.1, m.nnobsinode=10	7	0.6	0.53	0.59
Linear Discriminant Analysis			7	0.76	0.73	
K-Nearest Neighbor			7	0.43	0.34	

4.5 Model Deployment

Furthermore, with the aid of the DistilBert model and XGBoost algorithm, we deployed the model for Realtime accessibility and classification of infectious disease. Nevertheless, we present the final Application Programming Interface for the ecological infection disease classification where a user can describe the ecological factors within his or her environment and proceed to click the classifier button which then classifies the disease with a confidence interval among all the classes of the disease. Figure 16 below provides an Ecological Classification Interface where a user can easily describe his or her environment and click a button to classify which kind of infectious disease the ecological risk factors belong to.

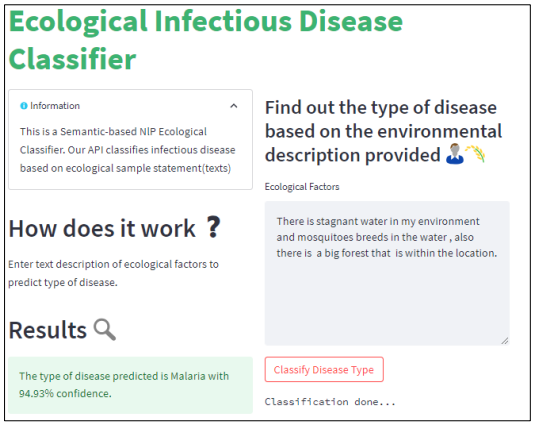

Figure 16: Ecological Classification Interface

Furthermore, we present the model interpretability results of the classification of infectious disease how the decision was made, and the criteria it uses in the decision making which is depicted in Figure 17.

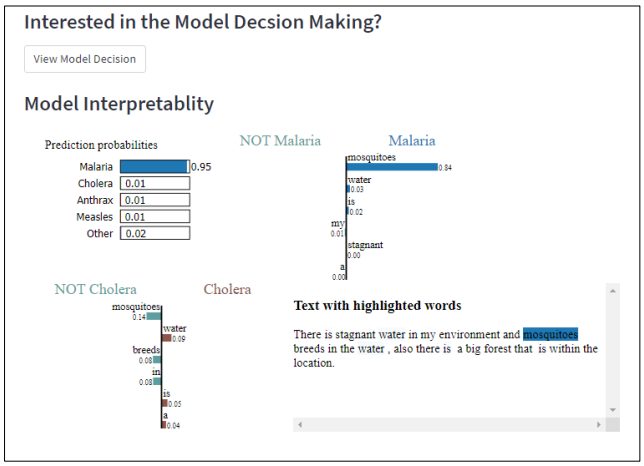

Figure 17: Model Decision Making

4. CONCLUSION

We deduce that this research unveils the intricate interplay of natural and human-induced factors in the development of infectious diseases, recognizing the pivotal role of the ecological aspect of human activities in disease transmission. The Framework-Based Method (FBM) offers a structured and reproducible approach to infectious disease classification, elevating the methodological rigor in the field. Notable variations in model performance highlight XGBoost and Linear Discriminant Analysis as superior, culminating in the integration of BERT with XGBoost to create an interactive interface. This integration not only enhances user experience and accessibility but also contributes significantly to the practical application of Machine Learning and Natural Language Processing (NLP) in ecological disease classification. Moreover, the research introduces a novel dimension by integrating model interpretability into ecological disease classification, addressing the challenge of transparency in advanced machine learning applications. This addition ensures that decision-making processes within the models are understandable and trustworthy, fostering confidence among stakeholders in infectious disease management. In the broader context, the research aligns with the hypothesis that a holistic understanding of infectious disease dynamics necessitates a comprehensive approach, considering both ecological and anthropogenic factors. The identified problem of incomplete methodologies in disease classification is mitigated through the FBM, providing a clear framework for data processing and model training. Ultimately, this research contributes to advancing the field of infectious disease management by not only enhancing classification accuracy and user accessibility but also by introducing a new paradigm of interpretability crucial for effective decision-making and public health strategies.

REFERENCES

[1] S. Morse, "Factors in the emergence of infectious diseases," in *Plagues and Politics*, A. T. Price-Smith (Ed.). Palgrave Macmillan, London, 2001, pp. 8-26. DOI: 10.1057/9780230524248_2
[2] S. Tong and C. L. Soskolne, "Global Environmental Change and Population Health: Progress and Challenges," *EcoHealth*, vol. 4, pp. 352-362, 2007.
[3] M. Sharma and A. Atri, *Essentials of International Health*, Jones & Bartlett Learning, 2010.
[4] H. Frumkin, "Urban sprawl and public health," *Public Health Reports*, vol. 117, no. 3, pp. 201-217, May-Jun. 2002.
[5] M. Naghavi, H. Wang, and R. Lozano, "Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990–2013: A systematic analysis for the Global Burden of Disease Study 2013," *The Lancet*, vol. 385, pp. 117-171, 2015.
[6] P. Grandjean and P. J. Landrigan, "Neurobehavioral effects of developmental toxicity," *The Lancet Neurology*, vol. 13, no. 3, pp. 330-338, Mar. 2014.
[7] K. Taniguchi et al., "Overview of infectious disease surveillance system in Japan, 1999-2005," *Journal of Epidemiology*, vol. 17, no. Suppl, pp. S3-S13, Dec. 2007. DOI: 10.2188/jea.17.s3
[8] K. E. Jones et al., "Global trends in emerging infectious diseases," *Nature*, vol. 451, no. 7181, pp. 990-993, Feb. 2008. DOI: 10.1038/nature06536
[9] S. L. LaDeau et al., "Data-model fusion to better understand emerging pathogens and improve infectious disease forecasting," *Ecological Applications*, vol. 21, no. 5, pp. 1443-1460, Jul. 2011. DOI: 10.1890/09-1409.1
[10] M. X. Tong et al., "Infectious diseases, urbanization and climate change: challenges in future China," *International Journal of Environmental Research and Public Health*, vol. 12, pp. 11025-11036, Sep. 2015. DOI: 10.3390/ijerph120911025

[11] H. Heesterbeek et al., "Modeling infectious disease dynamics in the complex landscape of global health," *Science*, vol. 347, no. 6227, p. aaa4339, Jan. 2015. DOI: 10.1126/science.aaa4339
[12] A. Wesolowski et al., "Connecting mobility to infectious diseases: The promise and limits of mobile phone data," *Journal of Infectious Diseases*, vol. 214, no. Suppl_4, pp. S414-S420, Dec. 2016. DOI: 10.1093/infdis/jiw273
[13] C. J. E. Metcalf et al., "Identifying climate drivers of infectious disease dynamics: Recent advances and challenges ahead," *Proceedings of the Royal Society B: Biological Sciences*, vol. 284, no. 1860, p. 20170901, 2017. DOI: 10.1098/rspb.2017.0901
[14] B. Davgasuren et al., "Evaluation of the trends in the incidence of infectious diseases using the syndromic surveillance system, early warning and response unit, Mongolia, from 2009 to 2017: A retrospective descriptive multi-year analytical study," *BMC Infectious Diseases*, vol. 19, no. 1, p. 705, 2019. DOI: 10.1186/s12879-019-4362-z
[15] M. Baguelin et al., "Tooling-up for infectious disease transmission modelling," *Epidemics*, vol. 32, p. 100395, Mar. 2020. DOI: 10.1016/j.epidem.2020.100395
[16] R. E. Baker et al., "Infectious disease in an era of global change," *Nature Reviews Microbiology*, vol. 20, no. 4, pp. 193-205, 2022. DOI: 10.1038/s41579-021-00639-z
[17] S. Inyang and I. Umoren, "Semantic-Based Natural Language Processing for Classification of Infectious Diseases Based on Ecological Factors," *International Journal of Innovative Research in Sciences and Engineering Studies (IJIRSES)*, vol. 3, no. 7, pp. 11-21, 2023.
[18] M. Muntean and F. D. Militaru, "Design Science Research Framework for Performance Analysis Using Machine Learning Techniques," *Electronics*, vol. 11, no. 16, p. 2504, Aug. 2022. DOI: 10.3390/electronics11162504
[19] I. A. Umoren et al., "A New Index for Intelligent Classification of Early Syndromic of Cardiovascular (CVD) Diseases Based on Electrocardiogram (ECG)," *European Journal of Computer Science and Information Technology*, vol. 11, no. 4, pp. 1-21, 2023.
[20] A. Ekong, A. Silas, & I. S. Inyang, "A Machine Learning Approach for Prediction of Students' Admissibility for Post-Secondary Education using Artificial Neural Network," *International Journal of Computer Applications*, vol. 184, pp. 44-49, 2022.
[21] I. J Umoren & S. J. Inyang, "Methodical Performance Modelling of Mobile Broadband Networks with Soft Computing Model," *International Journal of Computer Applications*, vol. 174, no. 25, pp. 7-21, 2021.