Comparative Analysis of Predictive Models for Diagnosis of Lower Respiratory Infections among Paediatric patients

[1]OLAYEMI, O. C., [2]OLASEHINDE O. O., [3]OJOKOH, B. A., [4]PETER, A. I.

[1]Department of Computer Science, Joseph Ayo Babalola University, IkejiArakeji, Nigeria.

[2]Department of Computer Science, Federal Polytechnic, Ile Oluji, Nigeria

[3]Department of Information Systems, Federal University of Technology, Akure, Nigeria.

[4]Department of Computer Science and Engineering, Obafemi Awolowo University, Ile-Ife, Nigeria.

ocolayemi@jabu.edu.ng[1], Olaolasehinde@fedpolel.edu.ng[2], baojokoh@futa.edu.ng[3], paidowu1@futa.edu.ng[4]

Abstract:

Lower Respiratory Tract Infections (LRTIs) are the major causes of mortality in paediatrics. Literature reviews reveal that LRTIs accounted for more than a million children morbidity and mortality yearly due to a lack of prompt diagnosis or no diagnosis due to a shortage of medical experts and medical facilities. The use of Machine learning (ML) techniques can be employed to fill this gap. This study evaluates ML models for a prompt and timely diagnosis of LRTIs in developing countries. The LRTIs dataset used in this study was obtained from The Federal Medical Centre, Owo, Nigeria. Relevant features of the dataset based on Information gain and Correlation feature selection techniques were used to build five machine learning predictive models; Naïve Bayes (NB), Multi-Layer Perceptron (MLP), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Random Forest (RF). The predictive models' evaluation was based on standard performance metrics (accuracy, sensitivity, specificity, and precision).The experimental results show that the information gain predictive models perform better than the correlation predictive models. The RF predictive model of the Information Gain feature selection method recorded the best accuracy of 98.53%. RF is therefore recommended for making decisions concerning respiratory infection diagnosis.

Keywords: Paediatric, Lower respiratory Tract Infection, Respiratory Rate, Cyanosis, Predictive Model**,**

1.0 Introduction

Respiratory conditions are the core causes of mortality in paediatrics in advanced and developing nations. They are accounted for more than 5 million deaths annually. Respiratory infections are the second most common cause of hospital-acquired infections [1]. In developed nations, respirational pollutions are one of the main causes of outpatient Consultations [2]. The human respiratory tract is made up of two divisions; The Upper and the Lower Tracts. The upper is positioned above the vocal folds and without signs of auscultation. The lower includes a whole series of situations that may or may not involve the parenchyma [3]. Pneumonia is lower tract infections that inflame the airbag in one or both lungs, sometimes filled with fluid. Bronchitis is the inflammation of the liner of bronchial tubes that carries the air to and from the lungs. All these infections can be triggered by bacteria, viruses, or even fungi [4]. Acute Respiratory Infections (ARIs) are a set of illnesses/indicators that constitute a principal source of paediatric illness and death of over 10 % of all children before their fifth birthdayinsub-Saharan Africa. Diagnosing respiratory infections turns tough when there are several other associated infections. Data mining is a valuable tool in the health sector and healthcare for discovering knowledge from a set of infection risk factors. Establishments that carry out data mining are well placed to meet their long-term needs. Patient's Data are great assets to healthcare establishments; Diagnosing and predicting a disease's outcome remains an exciting and inspiring task for data mining applications. Prediction is a supervised machine learning technique used to diagnose and predict a possible outcome based on the attributes of associated data instances [5].

Several studies have applied supervised machine learning algorithms to build prediction models in LRTI, Olayemi et al.[6] proposed a Naïve Bayes' classifier to predict the risk of LRTIs among paediatrics. The model recorded a diagnostic accuracy of 82%. Authors in [7]. Evaluate the predictive strengths of Naïve Bayes, Multi-layer perceptron, and K nearest neighbor classifiers in terms of diagnostic accuracy. This study's experimental results show that the Multi-layer perceptron classifier of the information gain reduced features of the LRTIs dataset recorded the highest diagnostics accuracy of 93.12%. A study in [8] shows that exposure to hydro-carbon and biomas from indoor pollution contributes to the risk of LRTIs among paediatric below 24 months. A hybrid model of Generic procedure and k-means clustering was developed in [9] to diagnose respiratory infections. The experimental result shows a significant improvement in the clustering accuracy when compared to ordinary K-means clustering. The Application of Machine Learning Techniques for LRTIs diagnosis was presented in [10].

From the reviewed papers, there is a need to have a comparative analysis of the diagnostic performances of the various machine learning models reviewed to determine models that will give optimal diagnosis performance of LRTIs. This paper is an extension of the work in [7]. It compares the predictive accuracies of five machine learning, namely: Support Vector Machine (SVM), Naive Bayes' (NB), Multilayer Perception (MLP), K-nearest neighbor (KNN), and Random Forest (RF), for the diagnoses of LRTIs in paediatrics, based on diagnostics accuracy and false diagnosis rate. It employs Information gain, and correlation feature selection algorithms were used to identify the relevant variables of the LRTIs dataset that have strong correlation/relevance to the diagnosis of LRTI. The identified relevant features were used to build predictive models for each of the five selected machine learning algorithms. The models were implemented using the R programming language.

2.0 Materials and Methods

The system architecture that describes the development of the model for diagnosing LRTIs infections in paediatric patients is shown in figure 1. The first stage is the Pre-processing stage, which involves discretizing the LRTIs datasets. The pre-processed datasets were split into 70% training and 30% testing. The second stage, the model building, is where all the datasets' features and the reduced features selected by the filtered-based feature selection Techniques (correlation-based technique and Information Gain Technique) were used to train the five-machine learning algorithms to build the LRTIs model. The third stage consists of the evaluating stage, where we built models using the whole testing feature sets and the reduced feature sets.

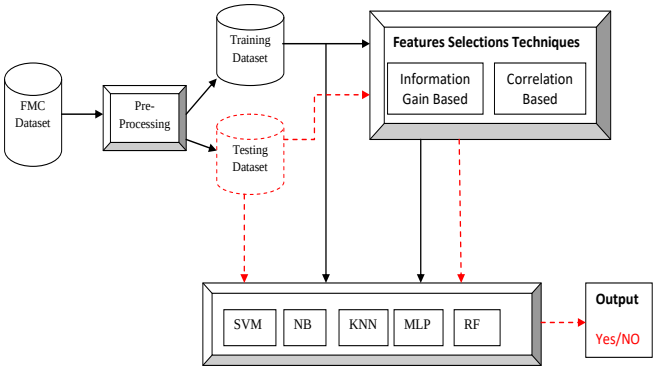

Figure 1: Architecture of the Predictive Models for Diagnosis of LRTIs

2.1 Data Collection and Pre-processing

A total of 1357 patients' records were collected from the Federal Medical Center (FMC) Owo, Nigeria, associated with the diagnosis and treatment of LRTIs. The datasets collected were 18 attributes with one class ID. The various input features identified for the diagnosis of Lower Respiratory Tract Infections in paediatric patients are as follows: gender, age, cyanosis, temperature, body mass index (BMI), difficulty in breathing, cough, fever. Respiratory rate, immunization, parents' smoking, parents' education level, exclusive breastfeeding, daycare, herbal mixture, overcrowding, and HIV. The

records were randomly split into two datasets: training and testing in a ratio of 70% and 30%, respectively. The dataset Cleaning and filtering of the dataset are necessarily carried out to avoid deceptive or inappropriate rules or patterns. Pre-processing of data to remove duplicate records, normalizing the values used to represent information in the database, accounting for missing data points, and removing unneeded data fields. The dataset was analyzed using a statistical univariate frequency distribution of all the dataset's features' discretized (nominal) values. Table 1 shows the frequency distribution of the initial features identified in the datasets. It also presents the percentage distribution of the values of each feature identified. A total dataset collected consists of (86%) records of Pneumonia patients, 4% records of Bronchitis patients, and (10%) of Bronchitis patients. The dataset was discretized since they were made up of discrete attribute values. The discretization was done before training commences on the dataset. The discretization technique reduces data size and uses class information, which may assist in improving classification accuracy. The total dataset was entered into an excel spreadsheet format for pre-processing (data cleaning, missing values, and incomplete instances). The extraction of relevant features was done to determine or select the relevant attributes to the target class. Filter-based feature selection methods (Information-based and correlation-based) was used to identify the most relevant features for the diagnosis of LRTIs. The two feature selection methods were chosen due to their simplicity and ability to select variables based on their identification of variables that have high relevance with determining the target class proposed by the physicians for each patient. The description of the dataset attributes is presented in Table 1, while the results are also shown in Table 2. The formulation of the predictive models was done using five different machine learning techniques. In this work, the attributes that were identified were directly correlated to LRTIs infections. Medical experts validated these attributes.

Table1: Distribution of the Identified Features in the Original Dataset

Types	**Variable Names**	**Attribute Values**
Input Variables	Gender	Male, Female
	Age (years)	Above1, Below one year
	Cyanosis	Yes, No
	Temperature	Abnormal, Normal
	BMI	Low, Normal, Very L
	Diff	Yes, No
	Cough	Yes, No
	Fever	Yes, No
	Resp. Rate	Yes, No
	Immunization	Yes, No
	Parents Smoking	Yes, No
	P. Education	Yes, No
	Ex. Breast F.	Yes, No
	Breast .F	Yes, No
	Daycare	Yes, No
	Herbal Mix	Yes, No
	Overcrowding	Yes, No
	HIV	Yes, No
Target Class	Diagnosis of LRTIs	Yes, No

2.2 Model Formulation

Supervised machine learning (SML) algorithms make it possible to assign a set of records X's (input variables of LRTIs) to a target class – the measurement of LRTIs (Yes or No) as seen in equation 1.

$$LRTI \quad LRTI = f(X)\begin{cases} yes \\ No \end{cases} \tag{1}$$

Five supervised machine learning algorithms were chosen for the formulation of the predictive models for the diagnosis of LRTIs in paediatrics patients of the southern part of Nigeria, namely: Naïve Bayes' (NB), Multi-layer Perceptron (MLP), K Nearest Neighbor (KNN), Support vector machines (SVM) and Random Forest (RF). The MLP and SVM algorithms also fall under perceptron network systems since input values are fired into nodes with synaptic weights assigned – inputs are the sum of products of weights $_{wi}$ and input xi, equation (2) shows the expression.

$$\sum_{k=1}^{i} w_k x_k = w_1 x_1 + w_2 x_2 + \cdots + w_i x_i =< w.x > \tag{2}$$

2.2.1 Support Vector Machines

A support vector machine is a technique that uses the representation of examples as points in space, mapped so that the examples of the separate categories are segmented by a clear gap that is very wide. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on. SVM constructs a hyper-plane or hyper-plane set in a high-dimensional space applied for classification, regression, or any other task. The hyperplane achieves a good separation with the largest distance to the nearest training data points *x* called the support vectors since, in general, the larger the margin, the lower the generalization error of the classifier.

Consequently, in formulating the SVM model, its attempts to minimize the cost by maximizing the distance between hyper-planes. A good separation is achieved by the hyperplane $< \boldsymbol{w}, \boldsymbol{x} > + b = 0$ that has the largest distance $\frac{2}{||w||}$ to the neighboring data points of either class at opposite ends since, in general, the higher the margin, the lower the generalization error of the SVM classifier. Figure 2 shows the separation of the different classes of LRTIs dataset.

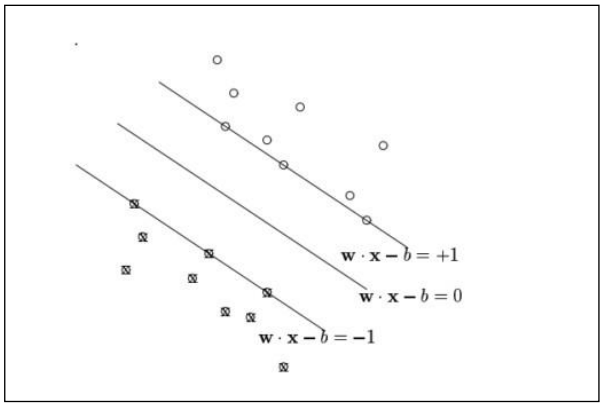

Figure 2: A linearly separable hyperplane using SVM

Assuming the dataset used in the study containing N training datasets $\{(x_1, y_1), (x_2, y_2)\ldots(x_{N,,,} y_N)\}$, where $x_i \in \mathbb{R}^d$ and $y_i \in \{\pm 1\}$. The motivation is to learn a linear separating hyperplane classifier:

$$f(x) = sgn < w.x - b > \tag{3}$$

Furthermore, there is a need for the hyperplane to have a maximum separating margin to the two classes. Specifically, there is a need to find the Hyperplane H; y = w; x – b = 0 and two hyperplanes, $H_i \; for \; i = 1,2$ parallel to it and with equal distance to it. (see figure 2) defined as in equation (4), while the problem can be formulated as equation 5.

$$y_i(w.x_i - b) \geq 1 \; for \; y_i = \pm 1 \tag{4}$$

$$\frac{min}{w,b}\frac{1}{2}w^T w \; subject \; to y_i(w.x_i - b) \geq 1 \qquad (5)$$

This equation is a convex, quadratic programming problem (in w,b) in a convex set. By introducing Lagrange multipliers $\propto_1, \propto_2, \ldots \propto_N \geq 0$, we have the following Lagrangian.

$$LL(w, b, \propto) \equiv \frac{1}{2}w^T w - \sum_{i=1}^{N} \propto_i y_i(w.x_i - b) + \sum_{i=1}^{N} \propto_i \qquad (6)$$

2.2.2 Naïve Bayes' Classifier

Naive Bayes' classifier is a probabilistic model that depends on Bayes' theorem. It is known as a statistical classifier. It is one of the habitually used methods for supervised learning. It provides a capable way of dealing with any number of attributes or classes, purely based on probabilistic theory. Bayesian classification provides practical learning algorithms and prior knowledge of observed data [11]. Let X_{ij} be a dataset sample containing records (or instances) of *I* number of risks factors (attributes/features) alongside their respective diagnosis of lower tract infection (LTI), *C* (target class) collected for *j* number of records/patients, and H_k = {H_1 =Yes, H_2 =No} be a hypothesis that X_{ij} belongs to class C. For the classification of the diagnosis of LRTI given the values of the jth record's risk factor, Naïve Bayes' classification required the determination of the following:

a. $PP(H_k|X_{ij})$ – Posteriori probability: is the probability that the hypothesis, Hk, holds given the observed data sample X_{ij} for $1 \leq k \leq 2$.
b. $P(H_k)$ - Prior probability: is the initial probability of the target class $1 \leq k \leq 2$;
c. $P(X_{ij})$ is the probability that the sample data is observed for each risk factor (or attribute), I; and
d. $P(|X_{ij}|H_k)$is the probability of observing the sample's attribute, X_i, given that the hypothesis holds in the training data X_{ij}.

Therefore, the posterior probability of a hypothesis H_k is defined according to Bayes' theorem, as shown in equation (2), while the LRTI class's determination is in the equation.

$$PP(H_k|X_{ij}) = \frac{\prod_{i=1}^{n} P(X_{ij}|H_k)P(X_i)}{P(H_k)} \quad for \; k = 1,2 \qquad (7)$$

2.2.3 K-Nearest Neighbor (KNN)

KNN can be described as learning by similarity , as it is learned by comparing a specific test tuple with a set of training tuples that are similar to it. It is classified based on the class of their closest neighbors. Most times, more than one neighbor is taken into consideration; hence, the name K-Nearest Neighbor (K-NN), the ''K'' indicates the number of neighbors taken into account in determining the class [12]. In this paper, our data tuples are restricted to patients with LRTIs symptoms as having Temperature, Fever, Cough, Cyanosis, low Weight, Incomplete Immunization, etc. The Euclidean distance between a training tuple and a test tuple can be derived as follows:

$$let \; p_i \; be \; an \; input \; tuple \; with \; p \; features \; of \; LRTIs \; (p_{i1}p_{i2}p_{i3})$$
$$let \; n \; be \; the \; total \; number \; of \; input \; tuples \; of \; LRTIs \; i = 1.2 \ldots., n)$$
$$let \; k \; be \; the \; total \; number \; of \; features \; of \; LRTLs \; (j = 1,2, \ldots, k)$$

The euclidean distance between tuple p_i and P_t (t = 1, 2,…,n) *can be defined as*:

$$dd(p_i p_t) = \sqrt{(p_{i1} - p_{t1})^2 + (p_{i2} - p_{t2})^2 + \cdots + (p_{in} - p_{tn})^2} \qquad (8)$$

$$in \; general \; term: the \; euclidean \; distance \; between \; two \; tuples, for \; instances, are$$

$$p_1 = (p_{11}, p_{12}, \ldots p_{1n}) and \; p_2 = (p_{21}, p_{22}, \ldots, p_{2n}) \; will \; then \; be:$$

$$distp_1p_2 = \sqrt{\sum_{i=1}^{n}(p_{1i} - p_{2i})^2} \qquad (9)$$

Equation (10) applies to the numeric attribute of LRTIs, in which we take the difference between each corresponding values of attributes tuple P_1 and P_2, square the result and add them together to get the square root of the accumulated result; this gives us the distance between the two points P_1 and P_2. From equation (9), the input instance diagnosis is based on the closest n neighbor.

2.2.4 **Multi-layer Perceptron (MLP)**

Looking at Artificial Neural Networks (ANNs), they are usually presented as systems of interconnected neurons that send messages to one another. Each connection has numeric weights that can be tuned based on familiarity, making neural nets adaptive to inputs, and capable of learning [12]. The meaning of the network is the interconnections between the neurons in the different layers of each system. The first layer has input neurons (LRTI prediction indicators) that send data via synapses to the middle layer of neurons and then via more synapses to the third layer of output neurons. The synapses store parameters are called weights that manipulate the data stored. In this study, the input neurons are represented by each LRTI prediction indicator variables determined by the following,

$xi = x_1, x_2, x_3, \ldots x_i\}$ where i is the number of variables (input neurons). The expression represents the effect of the synaptic weights, Wi, on each input neuron at layer j:

$$zj=_{w1jx1}+_{w2jx2}+\cdots\ldots_{w3jx3}+b- \quad (10)$$

For each neuron, $_j$ its output O_j is defined as:

$$Oj=(netj)=\varphi(\sum wijxink \quad (11)$$

The input net_j to a neuron is the weighted sum of outputs of the previous neurons. The number of input neurons is n, and the variable wij denotes the weight between neurons i and j. The backpropagation algorithm is used for training, and the mean square error between the actual and desired output is reduced to a predetermined level. The final weights are used to predict the risk of Paediatrics LRTIs.

2.2.5 **Random Forest (RF)**

Random forest is an ensemble of decision tree classification algorithm. Each of the decision trees is independent of each other in their predictions. RF uses bagging and feature randomness when building each tree to create an uncorrelated forest of trees whose committee's prediction is more accurate than that of any individual tree [13]. RF performs well with large datasets and can handle both binary and multi-class label problems.

2.3 **Performance Metrics**

To evaluate the performance of the supervised machine learning algorithms used for the classification of the diagnosis of LRTI, a confusion matrix of the classification of each predictive model was plotted.(Figure 3). The four parameters used to formulate the metrics are as follows:

a. True positives (TP) are correctly classified Yes cases;
b. False positives (FP) are incorrectly classified No cases;
c. True negatives (TN) are correctly classified No cases; and
d. False negatives (FN) are incorrectly classified Yes cases.

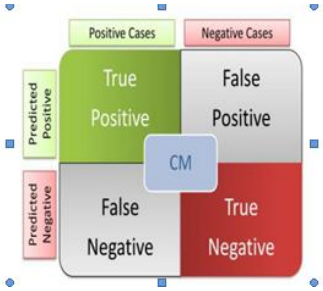

Figure 3: Diagram of a Confusion Matrix

The true positive/negative and false positive/negative values recorded from the confusion matrix can then evaluate the prediction model's performance. A description of the definition and expressions of the metrics are presented as follows:

a. True Positive (TP) rates (sensitivity/recall) – the proportion of positive cases correctly classified.

$$TP\ rate_{Yes} = \frac{TP}{TP + FN} - - - - - \quad (12)$$

b. $$TP\ rate_{No} = \frac{TN}{FP+TN} - - - - - \quad (13)$$

b. False Positive (FP) rates (1-specificity/false alarms) – the proportion of negative cases incorrectly classified as positives.

$$FP\ rate_{Yes} = \frac{FP}{FP + TN} - - - \quad (14)$$

$$FP\ rate_{No} = \frac{FN}{TP+FN} - - - - \quad (15)$$

c. Precision – the proportion of predicted positive/negative cases that are correct.

$$Precision_{Yes} = \frac{TP}{TP + FN} - - - \quad (16)$$

$$Precision_{No} = \frac{TN}{TN+FP} - - - \quad (17)$$

d. Accuracy – the proportion of the total correct predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} - - - - - - - - - - (18)$$

3.0. Results and Discussion

The resulting predictive models induced using the relevant features were compared based on the performance evaluation metrics with predictive models induced using the original features identified in the dataset. The predictive models' performance for the diagnosis of LRTI induced using each of the supervised machine learning models was evaluated to know the combination of feature selection and machine learning models suitable for developing an optimal predictive model for the diagnosis of LRTI. Table 2 shows the relevant features selected by each of the two filter base feature selection techniques used in this research. Following the application of the feature selection techniques for identifying the most relevant features, it was observed that cyanosis and temperature were the only features that were commonly selected by the feature selection algorithms. Thus, they were the features observed to be the most relevant based on the two methods of evaluating the relevance and having the greatest importance in the diagnosis of LRTIs among the identified patients.

The five techniques used in the classification of the predictive models of LRTI diagnosis for Paediatric patients are as observed in Table 3 are as follows: The table shows the confusion matrix of the evaluation of all the models used on the test dataset their performances. When all the eighteen (18) features identified from the datasets were used with the predictive models. The table shows that RF recorded the highest correctly classified 352 instances and 55 incorrectly classified instances with 86.49% accuracy. In comparison, KNN has the least accuracy in which correctly classified instances are 295 and 12 incorrectly classified instances with 72.48% accuracy. The correlation-based feature selection technique identified six (6) features that are directly correlated to the LRTIs. When these six features were used for the evaluation with the five techniques, It shows from the table that RF recorded the highest correctly classification of

400 instances and seven incorrectly classified instances with 98.28% accuracy. Similarly, KNN has the least accuracy of which correctly classified instances are 360 and 47 incorrectly classified instances with 88.45% accuracy in the group. The information-based feature selection technique identified ten (10) relevant features. It also shows from the table that RF recorded the highest correctly classification of 401 instances and six incorrectly classified instances with 98.53% accuracy.In comparison, KNN has the least accuracy of which correctly classified instances are 364 and 44 incorrectly classified instances with 89.43% accuracy in the group.It was observed that using the two reduced and relevant features of the LRTI for the diagnosis produce a more accurate result than using original features clinically identified by the experts.

The results revealed that the model performance was improved using the features selected via feature selection methods compared to the originally identified features. Figure 3 shows the graphical result of the entire performance. Random Forest algorithm attains 98.53% with ten features selected by the information-based feature selection method. So it can be deduced that the Information-based features selection method with Random forest is the best-performing algorithm on that particular group. The result further shows that RF is the most suitable in carrying out the diagnosis of LRTIs in paediatric patients and is recommended to build a diagnosis model.

Table 2: Relevant attributes identified using feature selection methods

Feature Selection Method	Information-Based	Correlation-Based
Search Method	***Ranker Search***	***Genetic Search***
Variables Selected	Age Sex Diff Cyan BMI Temperature Cough Fever Respiratory Rate Heart Rate	Cyanosis Temperature Coughing Imm Day Care HIV

Predictive Features	**Supervised Machine Learning**	**TP**	**FP**	**TN**	**FN**	**ACC (%)**	**FAR (%)**
Information based feature selection Technique (10 features)	NB	20	8	352	27	91.40	2.22
	MLP	24	4	362	17	94.84	1.09
	KNN	18	10	346	33	89.43	2.81
	SVP	23	5	356	23	93.12	1.39
	RF	26	2	375	4	98.53	0.53
Correlation-based Features	NB	18	10	345	34	89.19	2.82

Selection Technique (6 features)	MLP	23	5	358	21	93.61	1.38
	KNN	18	10	342	37	88.45	2.84
	SVM	22	6	349	30	91.15	1.69
	RF	26	2	374	5	98.28	0.53
Without Feature Selection Technique (18 Features)	NB	17	11	291	88	75.68	3.64
	MLP	20	8	302	77	79.12	2.58
	KNN	17	11	278	101	72.48	3.81
	SVM	19	9	299	80	78.13	2.92
	RF	23	5	329	50	86.49	1.50

Table 3: Comparative Analysis Results of all Predictive Models

4.0 Conclusions

In this paper, the comparative analysis of predictive models for diagnosing LRTIs among paediatrics patients was proposed using datasets from patients in FMC Owo, Ondo State, Nigeria. 18 variables were identified by the paediatrics Doctor to be necessary for predicting LRTIs in paediatrics patients, for which a dataset containing information of 1357 patients was provided with 14 attributes following the identification of the required variables. After data collection and pre-processing, five supervised machine learning techniques and two feature selection methods were used to develop the predictive model for the comparative analysis of LRTIs in Paediatric patients using historical datasets from which the training and testing were collected. The models were implemented using the R programming language, and also, the evaluation of the proposed models was carried out based on standard performance metrics (accuracy, sensitivity, specificity, and precision). The outcome results revealed that Random Forest (RF) with the Information Gain feature selection method proved to be an effective algorithm for diagnosing LRTIs among paediatric patients. Also,it is believed that higher accuracy could still be attained by using the Random forest method without the feature selection method. Rule induced algorithms can also be used to plot the relationship between the selected attributes identified to determine the comparative analysis of predictive models using the decision trees algorithm for further studies.

Acknowledgments

The authors acknowledge the following institutions and individuals for their support and contributions towards this research work; The Federal Medical Centre, Owo Ethical Committee, Nigeria. The members of staff of the Record Department, Federal Medical Centre, Owo, Nigeria. Dr. Bello, Head of Paediatrics Department, State Specialist Teaching Hospital, Akure, Nigeria Dr. Olaniyi Oluwole, Jobatec Clinic, Akure, Nigeria.

Ethical Standard Funding

This research work is self-funded research undertaken by the authorsat the Department of Computer Science, School of Computing, Federal University of Technology, Akure, Nigeria.

Conflict of Interest

The corresponding author states that there is no conflict of interest.

References

1. Yaya, S., and Bishwajit, G. (2019) Trends in the prevalence and care-seeking behavior for acute respiratory infections among Ugandan infants. glob health res policy 4(9). https://doi.org/10.1186/s41256-019-0100-8
2. Tam W.W., Wong T. W., Ng L, Wong S. Y., Kung K. K., and Wong A. H.(2014). Association between air pollution and general outpatient clinic consultations for upper respiratory tract infections in Hong Kong. PloS one. 9(1): e86913. DOI: 10.1371/journal.pone.0086913.
3. Jolien T., Berna D., and Katherine L (2016)'' Disease Course of Lower Respiratory Infection with a Bacterial cause'' Ann Fam Med 14(6): 534–539. doi: 10.1370/afm.1974
4. Graffelman A.W., Arie Knuistingh. N, Aloys CKroes .M, and Peter hans J. (2004).”Pathogens involved in Lower Respiratory Tract Infections in general practice" A British Journal of General Practice Br J Gen Practv.54 (498); PMC1314772.
5. Benko A., and Wilson B.(2003) "Online decision support gives plans an edge" Managed Healthcare Executive, 13(5): 20- 28.
6. OlayemiO.C., Adewale O. S, Olayemi O. O, Ojokoh B.A., and Adetunmbi A.O. (2018) ”Application of Machine learning to the Diagnosis of Lower Respiratory Tract Infection in Paediatric Patients” Paper presented at the 2nd International Conference on Information and Communication Technology and its Applications (ICTA)
7. Olayemi O. C. and Olasehinde O. O. (2019). Evaluation of Selected Machine Learning Techniques for Predicting Lower Respiratory Tract Infection in Paediatrics Patients. 3rd International Conference on Applied Information Technology (AIT2019)
8. P. A. Ahmed, K. K. Yusuf, A. Dawodu. (2015). Childhood acute lower respiratory tract infections in Northern Nigeria: At risk factors. Nigerian Journal of Paediatrics. 42 (3)
9. Jenn.L.Liu, Yu-Tzu .H, Chih-Lung .H. (2012)'' Development of Evolutionary Data Mining Algorithms and their Applications to Respiratory Disease Diagnosis'' WCCI 2012 IEEE World Congress on Computational Intelligence
10. [Ambali B.A, Ojo. A, andOladele .V A. (2017) Respiratory Infections among children prediction.'' JABU Journal of Science and Technology, 2(2): 23-30.
11. Hickey, Stephanie J. (2013) "Naive Bayes Classification of Public Health Data with Greedy Feature Selection," Communications of the IIMA: 13(2). http://scholarworks.lib.csusb.edu/ciima/vol13/iss2/7
12. Rupali .R. Patil (2014) '' Heart Disease Prediction System using Naïve Bayes'' Elsevier. Pp. 516–522. ISBN 978-1-4160-2973-1.
13. Alam Z, Rahman S, Rahman S. (2019). A random forest-based predictor for medical data classification using feature ranking. Inform Med 5:100180.https://doi.org/10.1016/j.imu.2019. 100180,

BIOGRAPHIES

Olayemi Olufunke. C is a Lecturer in the Department of Computer Science at Joseph Ayo Babalola University Ikeji-Arakeji, Osun State, Nigeria. She obtained her Ph.D in 2018 in Computer Science, Federal University of Technology Akure (FUTA), Nigeria.. Her Research/ Areas of Interest are Data Mining, Machine Learning, Bio-informatics and Artificial Intelligence.

Olasehinde Olayemi. O is a Lecturer in the Department of Computer Science at Federal Polytechnic Ile-Oluji, Ondo State, Nigeria. He obtained hisPh.D in 2019 in Computer Science from the Federal University of Technology, Akure, Nigeria. His Research/Areas of Interest are Cyber Security, Machine Learning, Text Mining, and Behavioural Analysis.

Ojokoh Bolanle. A is an Associate Professor in the Department of Information Systems at Federal University of Technology Akure, Nigeria. She had her Ph.D. in Computer Science, Federal University of Technology, Akure, Nigeria in 2009. Her Research/Areas of interest are Machine Learning, Data Mining and Knowledge Discovery, Text Mining, E-Learning and Digital Library

Peter Adebayo. I is an Associate Professor in the Department of Computer Science and Engineering, Obafemi Awolowo University, Nigeria. His research focus is on Applied Computing that is application of computing to address and solve health related problems in Sub Saharan Africa.