International Journal of Intelligent Information Systems
2020; 9(5): 44-55
http://www.sciencepublishinggroup.com/j/ijiis
doi: 10.11648/j.ijiis.20200905.11
ISSN: 2328-7675 (Print); ISSN: 2328-7683 (Online)

Stacked Ensemble Improvement of the Lower Respiratory Tract Infection Diagnoses in Peadiatric

Olasehinde Olayemi Oladimeji[1, *], **Olayemi Olufunke Catherine**[2], **Adetunmbi Adebayo Olusola**[3]

[1]Department of Computer Science, Federal Polytechnic, Ile Oluji, Nigeria
[2]Department of Computer Science, Joseph Ayo Babalola University, Ikeji-Arakeji, Nigeria
[3]Department of Computer Science, Federal University of Technology, Akure, Nigeria

Email address:
olaolasehinde@fedpolel.edu.ng (O. O. Olasehinde), ocolayemi@jabu.edu.ng (O. C. Olayemi), aoadetunmbi@futa.edu.ng (A. O. Adetunmbi)
[*]Corresponding author

To cite this article:
Olasehinde Olayemi Oladimeji, Olayemi Olufunke Catherine, Adetunmbi Adebayo Olusola. Stacked Ensemble Improvement of the Lower Respiratory Tract Infection Diagnoses in Peadiatric. *International Journal of Intelligent Information Systems*. Vol. 9, No. 5, 2020, pp. 44-55. doi: 10.11648/j.ijiis.20200905.11

Received: September 10, 2020; **Accepted**: October 5, 2020; **Published**: November 4, 2020

Abstract: Lower Respiratory Tract Infections (LRTIs) are the second and third causes of pediatric patients' death in Nigeria and the United States of America. It is observed from several reviewed literature that the LRTIs accounted for more than a million children morbidity and mortality yearly due to lack of prompt diagnosis or no diagnosis due to a shortage of medical experts and medical facilities in our localities. Intense research is ongoing on applying machine learning (ML) to its clinical diagnosis and reducing its spread in pediatric patients. In this research, K-Nearest Neighbor (KNN), C4.5 Decision Tree, and Naive Bayes' ML algorithms were used to develop three base diagnosis models with Correlation, consistency, and information gain selected feature of the LRTI dataset, Multiple Model Trees (MMT) Meta algorithm is used to combine and improve the diagnoses of all the base models using stacked ensemble. The preliminary diagnosis findings using base models have established that the information gained feature extraction method performed much better than the other two. It, therefore, suffix that the results from this should be used for further processing. All the models built with the reduced feature set recorded improved diagnoses accuracy more than the model built with the whole feature set. The MMT stacked ensemble models recorded an improvement on the diagnosis of LRTIs in Peadiatric, it recorded the highest diagnostic accuracies improvement of 12.80%, 13.52%, and 12.37%, and lowest diagnostic accuracies improvement of 6.37%, 5.22%, and 6.09% with the MMT stacked ensemble models of the Consistency, the Correlation, and the information gain reduced selected feature set respectively. These experimental results show the potential for this approach to deliver a reliable and improved diagnosis of LRTIs. It is recommended to be used to diagnose LRTIs in primary health care centers to reduce its mortality rate.

Keywords: Machine Learning Algorithm, Diagnosis, Stacked Ensemble, Infection, Diagnosis Accuracy, Incorrect Diagnosis Rate

1. Introduction

Diagnosis is the process of finding out the cause(s) of infection or patient sickness. It involves a physical examination, gathering information from the patient or caregiver, and laboratory tests. Correct diagnosis allows medical doctors and health givers to give correct treatment. The LRTIs are infections caused by bacteria, viruses, or fungi along the patients' respiratory tract [1]. Viral infections are the main causes of mild and moderate pneumonia (especially in the first year of a patient's life). Bacterial infections are the leading cause of severe pneumonia [2]. According to Worrall, Bronchioloties are the most common LRTIs in infants between 3 months to 6 months [3]. Bronchitis is a short term LRTIs of the airway affecting between 30 to 50 children in every 1000 children per year [4].

Respiratory diseases are responsible for mortality and morbidity among children all over the world. It is the second leading killer of children worldwide and accounted for about 20% of deaths in Pediatrics [5]. LRTI is responsible for the second leading causes of Pediatrics' death in Nigeria [6].

According to [7], LRTIs are the third-largest causes of death in the United States of America. Researchers have shown that LRTI can infect a child up to six times a year, and it accounts for 30% to 50% of all the pediatric outpatient visits [8]. Pneumonia is a primary cause of morbidity and mortality among children under five years in developing countries [9]. The risk of infections is on the increase due to many risk factors responsible for respiratory infections. Children cannot avoid having contact with the viruses and bacteria, which are the first risk factors in increasing their chances of developing respiratory infections. Children's regular contact with other children that could be infection carriers always puts them at risk. Also, most times, children often do not bother to wash their hands frequently and are also more likely to touch dirty things and put the hands in their mouths, resulting in the spread of these infections. Early detection and treatment of this infection will reduce its mortality rate.

Data mining (DM) is the computational process of extracting hidden knowledge from raw data volumes through algorithms and techniques drawn from statistics and machine learning. DM has many applications such as cyber-attacks detection, web mining, customer relationship management, medical analysis, and expert system.[10]. DM is a useful tool for clinical diagnosis of infections in primary health [11]. The application of data mining techniques in the clinical diagnosis of infections has proven to aid early diagnosis of infections in patients [12]. Machine learning (ML) is a DM technique used to extract or discover useful patterns in a dataset. This technique learns patterns of risk factors responsible for diseases or infections in many patients and uses it to predict a new patient's health status. ML has been widely used in the healthcare industry to diagnose various diseases, drug formulation, treatment recommendations, and disease outbreak and spread predictions [13]. Predicting the health status of a patient is an interesting application area of data mining. In this study, ML is used to build a diagnosis model that predicts health status as either infected or not infected with LRTI. The LRTI predictive model building involves the training of Machine Learning Algorithms with the LRTI training dataset and its evaluation with the LRTI test dataset.

Ensemble Learning (EL) is an ML technique that applies multiple learning algorithms to improve the predictive result [14]. The primary use of EL is to improve the prediction of predictive models. The stacked ensemble is an EL technique that employs a meta-algorithm, usually called second-level learners, to train and efficiently combine the output of multiple learners called the based learners. Multiple Model Tree (MMT) algorithm is a model tree induced with the M'5 algorithm. MMT algorithm builds a tree-based model with linear regression functions at its leaves. It can be used for the regression problem and classification problem by transforming it into a function approximation problem [20]. This paper applied Stacked Ensemble with Multiple Model Tree (MMT) as a meta-algorithm to improve the diagnoses of three LRTI based models built from three machine learning algorithms; k-nearest Neighbor Classifier (KNN), Naïve Bayes' (NB), and C4.5 Decision Tree.

2. Literature Review

The aim of the research of UmaMaheswari et al., was to predict and improve heart disease prediction using machine learning algorithms. Random Forest was used to selecting the Heart Disease Dataset's relevant features from the UCI machine learning repository. Twelve attributes out of the original 76 attributes of the dataset were selected and used to build the predictive base models of Random Forest, Generalized Boosted Regression Modeling, Linear Discriminant Analysis, and Support Vector Machine. The stacked ensemble was used to combine their base model predictions. The proposed method provides better accuracy when compared to other methods reviewed in the literature. [15]. Oguntimilehin et al., applied stacked ensemble to improve the diagnosis of malaria using stacked ensemble; two stacked generalization Meta-learning algorithms (Random Forest and NNGE) were used to combine the diagnoses of six base models (PART, REP Tree, J48, Random Tree, RIDOR and JRIP) with improving the diagnosis performance. The two results obtained were compared in terms of classification accuracy. The results show that NNGE as a Meta learner performs better than Random Forest [16]. Olayemi et al., presented a paper titled "the development of a predictive model for pediatric patients, with lower respiratory tract infection, using data mining approach. The paper made use of Naïve Bayes' classifier for predicting the risk of lower respiratory tract infections. The result shows that the model used was suitable for carrying out the predictive task with a minimum accuracy of 92% [17]. Jan et al., applied a Weighted Vote Ensemble technique to improve the diagnosis accuracies of five heterogeneous base models, using Cleveland and Hungarian UCI reposition heart datasets; Random forest (RF), Support vector machine (SVM), Neural Network, Naive Bayesian (NB), and regression analysis classification. The implementation of the ensemble models was carried out using the WEKA Data Mining package. Results from the experiment show that the ensemble models recorded superior diagnostic accuracy [18].

McDonough applied data mining and ensemble techniques to predict a high risk of Surgical Site Infections in Gynecologic Cancer Patients, seven (7) machine learning algorithms (NB, RF, KNN, SVM, Recursive Partitioning and Regression Trees (RPRT), Logistic Regression (LR), and Feed forward Neural Network (FFNN)) were used to build the base models. Three of the base algorithms were used as the Meta classified to improve the seven base models' predictions using a stacked ensemble. Out of the three Meta classifiers, SVM Meta classified recorded the best performance in terms of specificity, sensitivity, accuracy, and Area under the curve (AUC). The performance of the three Meta classifiers is better than all the base model performances [19]. Quinlan investigated classification via regression and reported that classification via Multiple Model Trees (MMT) performs extremely better than Multi Response

Linear Regression (MLR) and better than C5.0 (a successor of C4.5), MMT is a suitable choice for learning at the Meta-level, especially in domains with continuous attributes [20].

Olasehinde et al., applied ensemble learning to improve the network intrusion detection performance of three base leaners'. K Nearest Neighbor, 4.5 Decision Tree and Naïve Bayes, were used as base models to learn intrusive and normal network connection patterns using the UNSW NB15 Intrusion Detection Dataset. Their predictions served as input to the M5' induced Model Tree meta learner algorithm individually and are collectively evaluated using the ten-fold cross-validation to build the stacked ensemble model used for classifications of the network traffics into any of nine network attacks categories or normal packet traffics. The results from this research show that the MMT stacked model of the three base learners' predictions gives a higher multi-class classification accuracy than the best accuracy recorded by any of the three base models. It also recorded the highest classification accuracy of 97.93% and lowest false diagnosis rate of 0.22% for the binary (attacks and normal label) evaluation of the test dataset [21]. Kaveh et al., applied the M5' algorithm to develop a model tree for predicting the shear strength of the High Strength Concrete (HSC) slender reinforced concrete beams without a stirrup. The developed model recorded a superior performance than the most common design codes. The RMSE and R^2 values of the M5′ model show improvement by 37.8% and 60% compared to the AS 3600 model as the most precise model among the several design codes [22]. Duggal et al., investigated the performance of two Machine learning Algorithms; Decision Table Majority classifier and the conjunctive Rule Learner, with the M5-Rules algorithm for the modeling of Effort Estimation of Software Projects, Performances of these models are tested on NASA Software Project Data, and the results were compared with four other models mentioned in the literature. M'5 rule recorded the best performance with the lowest value of 377.5 and 801.09 for Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). M'5 rule also recorded the best estimation capability; hence, it is recommended to build a suitable software effort model structure [23].

3. Lower Respiratory Tract Infections (LTRIs) Dataset

LRTI clinical risk factors of the two thousand one hundred and ten (2110) paediatric patients used for this research were collected from the Federal Medical Centre (FCM) Owo, between November 2015 to October 2016. Purposive data collection technique was used, and the health institution authorities granted direct access to the medical records of clinically diagnosed patients to have respiratory infections. The collected patient's record contained eighteen (18) independent attributes and one (1) dependent attribute (class identity). One thousand, eight hundred and fifty-four (1854) instances of the collected patient's records were clinically diagnosed to have respiratory infections, which consist of one thousand, five hundred and twenty (81.98%) instances of Pneumonia patients, one hundred and eleven (5.99%) instances of Bronchitis patients and two hundred and twenty-three (12.03%) instances of Bronchiolitis patients, all these patient records were stored with class identity "infected" with LTRI. The remaining two hundred and fifty-six (256) collected patient's records were clinically diagnosed with other infections such as HIV, Tuberculosis, Typhoid Fever, Malaria Fever, and were stored with class identity; "Not Infected" with LTRI. The data were collected in a raw form and later stored in the MS-Excel Xls format as an LRTI diagnosis dataset, as reported in Table 1.

Table 1. *Distribution of LRTI Diagnosis Dataset Infected and Non infected Instances.*

	LRTI Infected	Non-LRTI Infected	Total
No of Instances	1854	256	2110

A detailed description of each of the attributes of the LRTI diagnosis dataset reported in Table 2, is as follows:

a) Sex: is a description of the gender, which is a nominal value of male or female.
b) Age is a description of the present age of the patient receiving treatment, and it is recorded as a numeric value which determines the age in months: that is, a child of 2 years will have an age of 24 months.
c) Weight: A description of patients' weight at the point of receiving medication: the weight is a numeric value measured in Kilograms (kg). The age and weight of a child altogether are required in determining the nutritional status of the child.
d) Breastfeeding (BF): lack of BF and weaning baby prematurely is associated with some health challenges to the mothers and their babies, babies that are well breastfed for at least four months and after that, are not likely to have respiratory and gastrointestinal risks. The value can either be yes or no.
e) Parental smoking: this is a description of parents or caregiver that smokes.
f) Cyanosis: This is an indication of a low level of oxygen in the red blood cells of the patient, a bluish cast on the palm and mucous membrane is an indication of Cyanosis.
g) Respiratory Rate (RR): This is the no of times a patient breath in a minute, usually measured when the patient is at rest; high/abnormal RR is an indication fever or illness
h) Cough: This is a common illness in pediatrics. It helps to clear the throat and the airway from foreign body aspiration, it is associated with cold, and it is one of the risk factors of LRTI.
i) Temperature: this measures the heat generated by in the body, Normal: 36.5–37.5°C (97.7–99.5°F), abnormal - Fever: >37.5 or 38.3°C (99.5 or 100.9°F)
j) Indoor air pollution exposure: This is an exposure to indoor air pollution (use of wood and biomass fuels) for

cooking.

k) Immunization: is the vaccination against some diseases in children. Lack of or incomplete Immunization for children can result in serious diseases in children.

l) Crowding: Crowding occurs when there are more than seven persons per bedroom (8 x 10 dimensions). Also, when more than four persons share a child's bedroom, according to (WHO 2012). The room is said to be *overcrowded*.

m) Fever: If a child has a high temperature or low body temperature, the child is likely to have a fever

n) HIV infection: This is when a child has tested positive to HIV/AID either from birth or after birth

o) Difficulty in breathing: Difficult breathing or shortness of breath, also called dyspnea, can be harmless due to exercise or nasal congestion. It may be a sign of a more serious heart infection.

p) Herbal mixture: Herbal mixtures are a combination of plant mix together. It may contain a whole plant, parts of a plant, or extracts of either one or a combination of plants mixed and given to a sick child or person like a medicine.

q) Educational Status: This is the level of the educational background of the parents or caregiver.

r) Daycare: This is daytime cares for people that cannot be fully independent, such as children or older people. Children would be looked after in daycare while mothers go to work.

s) Class Id: This is an indication of whether a patient is having LRTIs or not. It is the required output variable.

Table 2. *Attributes Description of Variables and Measurements of LRTIs.*

S/N	Abbreviation	Attributes	Attributes Type	Description
1	SEX	SEX	Discrete	1= Male/, 2= female
2	AGE	AGE	Discrete	Represented in months (1yrs =12, 2 yr =24)
3	BRFD	BREASTFEEDING FOR SUCKING BABIES	Discrete	1 =Yes, 0 = no
4	PASM	PARENTAL SMOKING	Discrete	1=Yes, 0 = No
5	CYAN	CYANOSIS	Discrete	1=Yes, 0 = No
6	COGH	COUGH	Discrete	1 =Yes, 0 = No
7	TEMP	TEMPERATURE	Discrete	1 =Yes (Normal < =36), 0 = No (High) >37)
8	RR	RESPIRATORY RATE	Discrete	1 =Yes (Normal), 0 = No (Abnormal)
9	HRRR	Heart RATE	Discrete	1 =Yes (Normal), 0 = No (Abnormal)
10	POLL	INDOOR AIR POLLUTION	Discrete	1=Yes, 0= No
11	IMMU	INCOMPLETE IMMUNIZATION	Discrete	1 = yes Complete immunization 0 = No Incomplete immunization
12	HIV	HIV/AIDS INFECTION	Discrete	1=Yes, 0= No
13	CROW	CROWDED	Discrete	1=Yes, 0= No
14	FEVE	FEVER	Discrete	1=Yes, 0= No
15	DIFF	DIFFICULTY IN BREATHING	Discrete	1=Yes, 0= No
16	PEDU	PARENT EDU. STATUS	Discrete	1=Yes, 0= No
17	DCAR	DAYCARE	Discrete	1=Yes, 0= No
18	WEGT	WEIGHT	Discrete	1=Yes, 0= No
	ID	CLASS IDENTIFICATION	Discrete	1=Yes indicate Infected, 0= No indicate Not Infected

All the dataset attributes were discretized to make it suitable for the machine learning algorithms to build the diagnosis models, as reported in Table 2. All the attributes with a possible value of "Yes" were discretized as the discrete value of one (1), and attributes with a value of "No" were discretized as discrete value zero (0). Value "No" implies that the symptom is not present in the patients. The first attribute, "Sex," the male is discretized as the discrete value one (1), while the female is discretized as discrete value two (2), for the attribute "Temperature" the temperature below 37.5°C is considered normal and discretized as discrete value (one) 1, while temperature more than 37.5°C is considered high, and discrete as discrete value zero (0). There are two values for the Class id attribute, Yes (1) means the LRTI infection is present in the patients while No (0) means the LRTI infection is not present in the patient.

4. Methodology

Figure 1 shows the architecture of the proposed Stacked Ensemble Accuracy Improvement of Lower Respiratory Tract Infection using the MMT Meta Algorithm; it consists of two sections, section one is the stacked ensemble model building section, it is indicated with continues the black arrow lines in figure 1, it consisted of three different stages; the dataset discretization stage handles the discretization of the LRTIs training dataset, to make it suitable for the machine learning base algorithms learning.

The second stage involves selecting the relevant features/attributes of the LRTIs training dataset used to build and evaluate the base models. Each of the selected features of the LRTI's dataset by the three features selection techniques employed in this work, and all the feature set were used to train the K Nearest Neighbor, C4.5 Decision Tree, and Naïve Bayes' base algorithms to build and evaluate the base diagnosis models via ten folds cross-validation evaluation technique separately.

In the last stage of the first section, the base models' diagnosis built from each of the reduced feature set, with the actual corresponding class id of each diagnosed instance, formed the feature Meta Level datasets. Each of the Meta level datasets was used for the second level training of the MMT Meta Algorithm separately to build MMT stacked ensemble models for each of the reduced feature set.

The second section is indicated with the short red dashed arrow lines in figure 1, each discretized and selected reduced features/attributes of the LRTI's test dataset were used to evaluate the three base models built in the second stage of section one. The test dataset diagnosis was used to evaluate the stacked ensemble model built in the last stage of section one to obtain an improved patient diagnosis as either LRTIs infected or not infected. The system was implemented using Python programming language on a Corel i3, 64bits, 2.4 GHz processor, 16MB Cache, 512GB SDD, Ms. Windows 7 operating system.

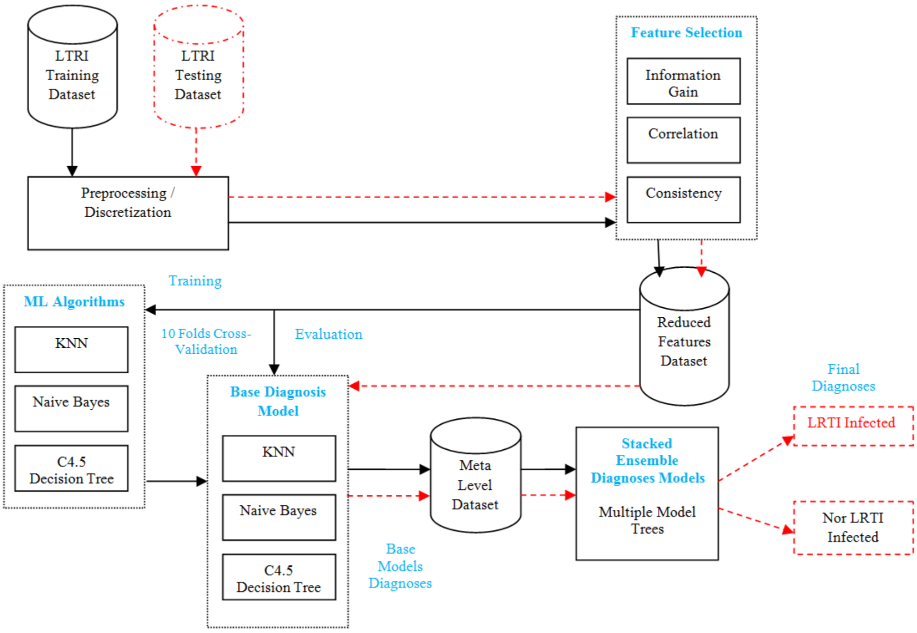


Figure 1. *Architecture of the LRTI Stacked Ensemble Diagnosis System.*

4.1. Feature Selection

Feature Selection (FS) is an important task in machine learning and model building; it can significantly improve predictive models' accuracy. It removes redundant attributes and selects relevant feature set for the ML algorithm training and model building. [24, 25]. According to Olayemi et al., Identification of the variables relevant to LRTIs diagnosis will likely improve the performance of the machine learning diagnosis models built with the selected feature set and also reduces the model's complexity [17]. This study employed three FS techniques: Consistency, Correlation, and Information Gain to select relevant features of the LRTI diagnosis dataset used in this study for the diagnosis models building.

Consistency Based Features Selection Techniques

Consistency Based Feature Selection (CBFS) generates all possible feature subset S of the LRTI dataset. For each of the generated subsets S, it computes the inconsistency count $\mathrm{INC_i}$ of all pattern $\mathrm{p_i}$ of the subset S and then calculates the inconsistency rate INCR of subset S using Equation 1. Subset with the lowest inconsistency rate INCR, it selects as the reduced feature set.

$$\mathrm{INCR} = \frac{\sum_{i}^{h} \mathrm{INC_i}}{\mathrm{M}} \qquad (1)$$

Where h is the number of all possible patterns of subset s of the LRTI dataset and M: is the number attributes (features) contained in the subset S of the LRTI dataset.

Correlation Features Selection Techniques

Correlation Features Selection (CFS) generates all possible attributes (features) subset S of the LRTI dataset, then used Equation 2 to calculates the $\mathrm{Merit_s}$ function of each subset S. The subset S that recorded the highest $\mathrm{Merit_s}$ value is selected as the reduced feature set.

$$Merit_s = \frac{k\overline{r_{cf}}}{\sqrt{k+k(k-1)\overline{r_{ff}}}} \qquad (2)$$

Where; K is the number of features/ attributes contained in the subset S

$\overline{\mathrm{r_{cf}}}$ = average feature-class correlation

$\overline{\mathrm{r_{ff}}}$ = average feature-feature correlation

Information Gain Selection Techniques

Information Gain Selection Techniques (IGFS) is a feature ranker. It computes the Information Gain (IG) using equation 3 for each LRTI dataset feature. It ranks each of the features based on the computed value of their IG in descending order, it then validates, set of the ranked features in terms of their diagnosis accuracy; on the training dataset, the set of ranked features that give the highest diagnosis accuracy is selected as the reduced feature set. Given an independent features X (LRTIs attributes) and dependent feature Y (Class Identity), the IG for the feature X is given in equation 3

$$\mathrm{IG(X)= H(Y)- H\,(Y\,|\,X)} \qquad (3)$$

Where H(y) is the Entropy of Y given in equation 4 and H(Y | X) is the Entropy of Y given X, given in equation 5

$$H(Y) = -\sum_{i=1}^{n} p(y_i) log_2 p(y_i) \qquad (4)$$

$$H(Y \mid X) = -\sum_{i=1}^{n} p(x_i) \sum_{j=i}^{k} p(y_i \mid x_i)\, log_2 p(y_i \mid x_i) \quad (5)$$

Where n: is number of instances in the LRTI dataset, k: is the number of possible class id values in the LRTI dataset, P (y_i): is the probability of occurrence of target class value of instance i

4.2. Base Models

Three machine learning algorithms, K Nearest Neighbor (KNN), Naive Bayes, and C4.5 Decision Tree, were adopted to build the base models.

K Nearest Neighbor

KNN is a supervised ML algorithm used for the regression and classification of predictive problems. It is a simple and easy to implement algorithm. A study in [26] reported that the KNN model recorded better classification performance when compared with other ML models. It used the Euclidean distance function to calculate the distance between the LRTIs test dataset instance being query and the rest of the LRTIs training instances to determine the K- nearest (closet) neighbor instance. It then classifies the queried instance as the class id (infected or not infected) with majority votes among the K chosen closest instances. The Euclidean distance function is given in equation 6.

$$d(p_i, \mathrm{q}_{it}) = \sqrt{\sum_{i=1}^{n} (p_i - q_{it})^2} \text{ (t = 1,...k)} \quad (6)$$

Where p_i is the instance of the LRTIs test dataset being queried, q_{it} is the training dataset instances. n is the number of features of both LRTIs training and test dataset, k is the number of instances in the training dataset. d(p_i, q_{it}) is the distance between p_i and q_{it}

From equation (6), a given instance will be classified as the class id (infected or not infected) with majority votes among the K chosen closest instances.

Naive Bayes

Naive Bayes (NB) is a simple and popular machine learning algorithm. It is very fast at predicting the class of a new instance and performs well in multi-class diagnosis; the study in [27] reported that NB is good at solving diagnostic and predictive problems. NB finds the probability of a class id where risk factors (attributes) relating to the class id are used as the input. It assumes that the risk factors of the LRTI dataset are independent of each other, it predicts the class id of a patient given its risk factors; it calculates the probability of each attribute, with each class id, uses the product rule to obtain a joint conditional probability of all the patient risk factors, then uses Bayes rule to derive conditional probability for each class label and predicts the class label of the given patient as the class label with the highest probability. The probability that a class label y_j will be assigned to a given unlabelled instance X of the LRTI dataset is given in Equation 7.

$$p(y_j \mid x_1, \ldots., x_k) = \frac{p(y_j)p(x_i|y_j)}{p(x_i)} \;(\forall_j = 0,1) \quad (7)$$

Maximum posterior probability for classifying a new instance as a class label is given in Equation 8

$$y = \underset{y}{\arg max}\; py_j \prod_{j=0}^{1} p\,(y_j)p\,(x_1, x_2, \ldots x_k \mid y_j) \quad (8)$$

Where $(x_1, x_2, \ldots, x_k)$ attributes of the LRTI dataset are called the predictors. y_j is the possible class values (infected, or nor infected). $p(y_j \mid x_1, \ldots., x_k)$ is the probability of class value yj given set of attribute values $(x_1, x_2, \ldots x_k)$. K is the number of the selected attributes by each feature selection techniques

C4.5 Decision Tree

C4.5 Decision Tree (DT) is a universally used classification technique for infection diagnosis. The study in [28] reported that the C4.5 classifier offers better performance and provides the pregnant woman a specific level of risk of a safe and healthy pregnancy period. C4.5 DT builds a diagnosis tree consisting of nodes, which are the attributes of the LRTI dataset, and arcs, which are attribute values. The arc connects to other nodes to the tree leaves, which are the class id (class label). The tree is used to predict the class id of a new patient. C4.5 DT calculates the Gain Ratio of all the attributes (X) of the training dataset using equation 9, and the Split value using equation 10. The attribute with the highest Gain Ratio is used to divide the dataset features into two subsets. The attribute with the highest gain ration of the two subsets is further used to divide each of the subsets to another two subsets; this procedure continues until a leaf node is reached, the patient will be diagnosed as the value of leaf node (class label) of the nodes that correspond to the risk factors attribute values.

$$\text{Gain Ratio(X)} = \frac{InformationGain(X)}{Splitinformation(X)} \quad (9)$$

$$Split(X) = -\sum_{x \epsilon X} \frac{|x|}{|n|} \cdot \log_2 \frac{|x|}{|n|} \quad (10)$$

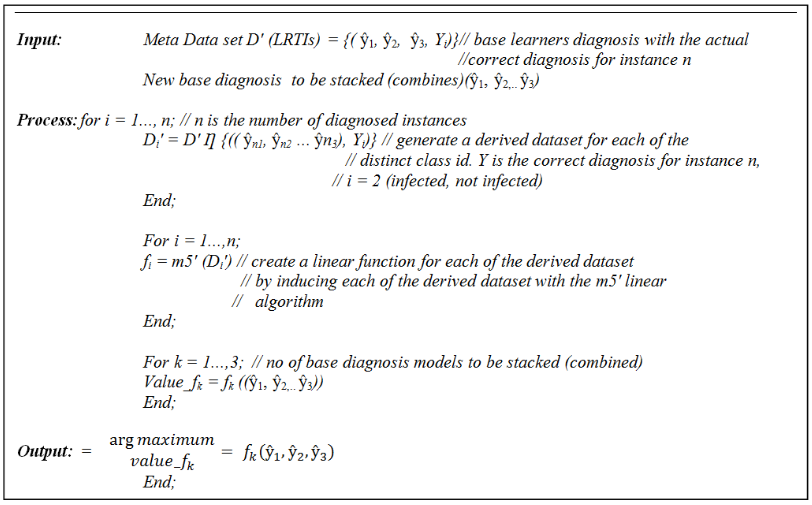
Where n is the number of values in attribute X and|x| is the value of the attribute X

4.3. Stacking with Multiple Model Trees (MMT)

The stacked framework consists of two phases; In the first phase, the LRTIs dataset S, consisting of instances of the form $s_i = (x_i, y_i)$ where x_i is feature vector, and y_i is the class id, was used to train machine learning algorithms L_1,.., L_3, using leave-one-out validation technique, to create base classifiers C_1, C_2,....., C_3, where $C_i = L_i(S)$, with the base models diagnoses results of the form $\hat{y}_1$, $\hat{y}_{2,..}$ $\hat{y}_3$. In the second phase, the base models' diagnosis results were used to train the Meta learner algorithm to build the MMT Meta learner model.

MMT, an M5' induced model trees, is a Meta learner adaptation of C4.5 decision tree with linear regression functions at the leaves. It can be adapted to diagnose and apply to classification problems by employing a standard method of transforming a classification problem into a function approximation problem [29]. Using this simple transformation, the model tree inducer M5’ generates more accurate classifiers than the state of the art C5.0 Decision Tree learner, particularly when most of the attributes are numeric.

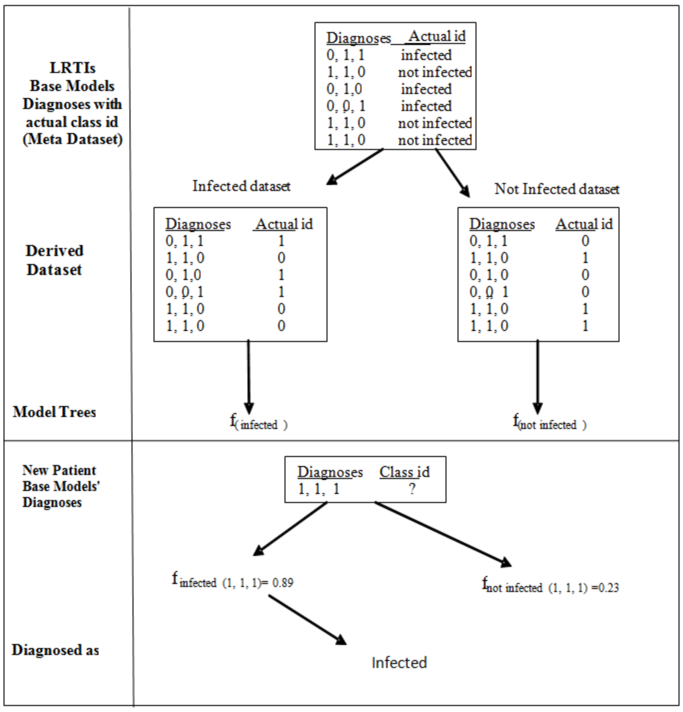
The algorithm of the M5' is presented in Figure 2, Meta Dataset D of the form {(X_1, Y_1), (X_2, Y_2),...., (X_n, Y_n)}, *where* (x_{i1}, x_{i2},....x_{i3})∈ X_i are the predictions of the three base models, for instance i of the LRTI dataset, (y_{i1}, y_{i2}) are the corresponding class id for each of the diagnosed instance i of the Meta dataset, and n is the number instances diagnosed. A derived dataset D_i', was created for each of the two possible class id (Infected, not infected). A linear regression function, f(D_i''), is generated for each of the derived datasets by inducing M'5 algorithm (M) on each of the derived datasets. To improve the base diagnosis of an instance K of the form (x_1, x_2,....x_3, function f_k for each the derived dataset is evaluated, instance K is classified as the derived dataset with the highest value of function k.



```
Input:          Meta Data set D' (LRTIs) = {( ŷ1, ŷ2, ŷ3, Yi)}// base learners diagnosis with the actual
                                                   //correct diagnosis for instance n
                New base diagnosis  to be stacked (combines)(ŷ1, ŷ2,.. ŷ3)

Process: for i = 1..., n; // n is the number of diagnosed instances
                Di' = D' I] {(( ŷn1, ŷn2 ... ŷn3), Yi)} // generate a derived dataset for each of the
                                          // distinct class id. Y is the correct diagnosis for instance n,
                                        // i = 2 (infected, not infected)
                End;

                For i = 1...,n;
                fi = m5' (Di') // create a linear function for each of the derived dataset
                            // by inducing each of the derived dataset with the m5' linear
                           //   algorithm
                End;

                For k = 1...,3;  // no of base diagnosis models to be stacked (combined)
                Value_fk = fk ((ŷ1, ŷ2,.. ŷ3))
                End;

Output: =   arg maximum  = fk(ŷ1, ŷ2, ŷ3)
              value_fk
                End;
```

Figure 2. *Algorithm of M5' Induced Model Trees.*

Figure 3 reports, MMT stacking procedure. It first generates a derived dataset for each of the class id and built regression trees for each of the derived dataset, each of the regression trees was induced with M5'MT algorithm to generate, diagnosis improvement function for each class id. The base model diagnoses of a new patient to be improved are plugged into each of the linear regression function, the function with the highest value is returned, and the patients will be diagnosed as the class id of the function with the highest value.


LRTIs Base Models Diagnoses with actual class id (Meta Dataset)
Derived Dataset
Model Trees
New Patient Base Models' Diagnoses
Diagnosed as
Infected dataset
Not Infected dataset
f(infected)
f(not infected)
f infected (1, 1, 1)= 0.89
f not infected (1, 1, 1) =0.23
Infected

Diagnoses	Actual id
0, 1, 1	infected
1, 1, 0	not infected
0, 1,0	infected
0, 0, 1	infected
1, 1, 0	not infected
1, 1, 0	not infected

Diagnoses	Actual id
0, 1, 1	1
1, 1, 0	0
0, 1,0	1
0, 0, 1	1
1, 1, 0	0
1, 1, 0	0

Diagnoses	Actual id
0, 1, 1	0
1, 1, 0	1
0, 1, 0	0
0, 0 1	0
1, 1, 0	1
1, 1, 0	1

Diagnoses	Class id
1, 1, 1	?

Figure 3. *The operations of the MMT Stacking Procedure (adopted [29]).*

4.4. Diagnosis Performance Metrics

Evaluation of the machine learning algorithm is an essential part of the predictive or diagnosis model. A confusion matrix is one of the most intuitive and easiest metrics used for finding the correctness and accuracy of the model. It has four possible outcomes, which are; True Positive (TP, correct positive diagnosis), True Negative (TN, correct negative diagnosis), False Positive (FP, incorrect positive diagnosis), and false negative FN, (incorrect negative diagnosis). Diagnosis Accuracy, False (wrong/Incorrect) diagnosis Rate, and Performance Improvement are the three metrics used to evaluate the diagnosis models' performance in this work.

Diagnosis Accuracy

Accuracy (ACC) is the ratio of all correct diagnoses to the total number of instances in the LRTI test dataset. It is given by Equation 11. An accuracy of 1 implies an error rate of 0

$$ACC = \frac{TP+TN}{FN+FP+FN+TP} \tag{11}$$

False Diagnosis Rate (FDR)

False Diagnosis Rate (FDR) or False Alarm Rate (FAR) is a measure that tells us what proportion of patients that we diagnosed as having LRTI had LRTI. It is given by Equation 12

$$FDR = FAR = \frac{TP}{TP+FP} \tag{12}$$

Performance Improvement

Performance Improvement (PI) is the ratio of changes in performance to the initial performance. It is given in Equation 13

$$\mathrm{PI} = \frac{\mathrm{FinalPerformance} - \mathrm{InitiatialPerformance}}{\mathrm{InitialPerformance}} \tag{13}$$

5. Results and Discussions

Table 3 reports the frequency distribution of the initial features identified in the dataset consisting of 2110 patients' records. It also presents the percentage distribution of the values of each feature identified. From the result, the majority of the patients were male, with a proportion of 66.39%, implying a ratio 2 to 1 for the male to female infection of LRTI. The age distribution shows that majority of the patients were below one year with a proportion of 84.84% of patients. The results also show that most of the patients have difficulty breathing, which is 54.78% and, the majority of the patients do not have Cyanosis with a proportion of 83.70%. It further shows that most patients have normal body weight with a proportion of 82.99% and the remaining with below normal weight, while the majority of the patients treated had abnormal temperatures with a proportion of 52.18%. The results also show that majority of the patients had cough represented by a proportion of 89.19%, 80.38% had a fever.

The majority of the patients also had an abnormal respiratory rate (RR) represented by a proportion of 72.60%, while 69.48% were also observed to have been exposed to pollution. The results also show that an almost equal number of patients were observed for those with and without Immunization and whose parents were either smokers or educated. The majority of the patients were not breastfed since a higher percentage falls below one year, and this is represented by a proportion of 58.91%, while about 58.58% attended daycare centers. The results further show that 62.8% of the patients were administered with the herbal mixture, while most of the patients were also observed to live in overcrowded environments (7 or more people in a room of 8 by 10) represented by a proportion of 66.21%. The results, however, show that majority of the patients were HIV negative.

Table 3. *Distribution of the Identified Features in the LRTI Dataset.*

Features	Labels	Frequency	Percentage (%)
Sex	Male	1401	66.39
	Female	709	33.61
Age	Above 1	320	15.16
	Below 1	1790	84.84
Difficulty in Breathing	Yes	1156	54.78
	No	954	45.22
Cyanosis	Yes	344	16.30
	No	1766	83.70
Weight	Low	156	7.39
	Normal	1751	82.99
	Very Low	203	9.62
Temperature	Abnormal	1101	52.18
	Normal	1009	47.82
Cough	No	228	10.81
	Yes	1982	89.19
Fever	No	414	19.62
	Yes	1696	80.38
Respiratory Rate (RR)	Abnormal	1532	72.60
	Normal	578	27.40
Pollution	No	644	30.52
	Yes	1466	69.48
Immunization	No	935	44.31
	Yes	1175	55.69
Parents Smoking	No	1173	55.59
	Yes	937	44.41
PEdu	No	1129	53.51
	Yes	981	46.49
Breast Feeding	No	1243	58.91
	Yes	867	41.09
Crowding	No	713	33.79
	Yes	1397	66.21
Day Care	No	874	41.42
	Yes	1236	58.58
Herbal mixture	No	785	37.20
	Yes	1325	62.80
HIV	No	2042	96.77
	Yes	68	3,23
Class Id	No	256	12.13
	Yes	1854	87.87

Table 4 reports the result of the feature selection techniques used in this study; the CFS approach selected twelve (12) relevant features, the IGFS results shows that there were ten (10) relevant features, the CBFS algorithm was used to select the subset of features highly correlated with the target class (LRTI) but with low or no Correlation with other features. This technique identified six (6) relevant features.

Table 4. *Relevant Attributes Identified using Three (3) Feature Selection Methods.*

Feature Selection Method	Consistency-Based	Information-Based	Correlation-Based
Variables Selected	Age Sex Diff Cyanosis Weight Temperature Poll Imm Parents Smoking PEdu HRR Breast F	Age Sex Diff Cyanosis Weight Temperature Cough Fever Respiratory Rate Heart Rate	Cyanosis Temperature Coughing Imm Diff HIV

Table 5 reports the confusion matrix, the diagnoses accuracy, and false diagnoses rate of the base models, and the ensemble models diagnoses, for each of the three models of the reduced feature set and the whole feature attributes. C4.5 DT base models recorded the highest diagnostics accuracy with the three reduced feature set; 91.63% with Consistency reduced feature set, 90.84% with Correlation reduced feature set, and 93.36% with information gain reduced set, closely followed by Naive Bayes base models; 90.05% with Consistency reduced feature set, 87.52% with Correlation reduced feature set and 91.31% with information gain reduced set, KNN based models recorded the least diagnosis accuracy of 86.41% with Consistency reduced feature set, 84.20% with Correlation reduced feature set and 87.99% with information gain reduced set among the base models. Based diagnosis models of Information Gain reduced feature set recorded the highest diagnosis accuracy and lowest incorrect diagnosis rate, followed by Consistency reduced feature set, while Correlation reduced feature set recorded the least diagnosis, ahead of the whole set models.

The MMT ensemble of the base models diagnosis recorded the highest diagnostic accuracy 97.47%, 95.58%, 99.05%, and 90.68% for the Consistency, the Correlation, the information gain reduced selected features, and all feature set respectively. It also recorded the lowest false rate of 1.48%, 2.24%, 0.37%, and 3.24% for the Consistency, the Correlation, the information gain, and all feature sets. The MMT stacking of the information gain base models recorded the best-stacked ensemble performances with the highest diagnostic accuracy of 99.05% and the least false alarm rate of 0.37%, closely followed by the MMT stacking of the consistency base models with a diagnostic accuracy of 97.47% and false alarm rate of 1.48%. The stacking of correlation base models' performance recorded the least diagnostic accuracy of 95.58% and a false alarm rate of 2.24%.

Table 5. *Diagnosis Models Confusion Matrix and Performances.*

Features Used to Build the Models	Diagnosis Models	Confusion Matrix		Diagnosis Accuracy %	False Diagnosis Rate %
Consistency Reduced Features	C4.5 Decision Tree	TP = 78 FP = 15	FN = 38 TN = 502	91.63	2.90
	Naive Bayes	TP = 75 FP = 18	FN = 45 TN = 495	90.05	3.51
	K- Nearest Neighbour	TP = 69 FP = 24	FN = 62 TN = 478	86.41	4.78
	Multiple Model Trees Stacked Ensemble	TP = 85 FP = 8	FN = 8 TN = 532	97.47	1.48
Correlation Reduced Features	C4.5 Decision Tree	TP = 78 FP = 15	FN = 43 TN = 497	90.84	2.93
	Naive Bayes	TP = 73 FP = 20	FN = 59 TN = 481	87.52	3.99
	K- Nearest Neighbour	TP = 63 FP = 30	FN = 70 TN = 470	84.20	6.00
	Multiple Model Trees Stacked Ensemble	TP = 81 FP = 12	FN = 16 TN = 524	95.58	2.24
Information Reduced Features	C4.5 Decision Tree	TP = 79 FP = 14	FN = 28 TN = 512	93.36	2.66
	Naive Bayes	TP = 77 FP = 16	FN = 39 TN = 501	91.31	3.09
	K- Nearest Neighbour	TP = 70 FP = 23	FN = 53 TN = 487	87.99	4.51
	Multiple Model Trees Stacked Ensemble	TP = 91 FP = 2	FN = 04 TN = 536	99.05	0.37
Whole Features	C4.5 Decision Tree	TP = 65 FP = 28	FN = 87 TN = 453	81.83	5.82
	Naive Bayes	TP = 59 FP = 34	FN = 101 TN = 439	78.67	7.19
	K- Nearest Neighbour	TP = 58 FP = 35	FN = 113 TN = 427	76.62	7.58
	Multiple Model Trees Stacked Ensemble	TP = 73 FP = 20	FN = 39 TN = 501	90.68	3.84

Table 6 reports the performance improvements of the MMT stacked ensemble model diagnoses over the diagnoses of all the base models in terms of diagnoses accuracy and false (incorrect) diagnoses rate, the MMT model recorded the highest diagnoses accuracy improvement `of 12.80%, 13.52%, and 12.37% on KNN models with Consistency, Correlation, and information gain reduced features respectively, it also recorded lowest diagnosis improvement with the C4.5 Decision Tree models of 6.37%, 5.22% and 6.09% for Consistency, Correlation, and information gain reduced features.

The MMT stacked ensemble model recorded the highest false alarm rate improvement on KNN models; 62.67% with Correlation reduced features, 69.04% with Consistency reduced features, and 91.80% with information gain. It also recorded the lowest false alarm rate improvement on C4.5 Decision Tree models; 23.55% with Correlation reduced features, 48.97% with Consistency reduced features, and 86.09% with information gain reduced features.

Table 6. *Diagnoses Accuracy and False (incorrect) Diagnoses Alarm Rate-Improvement of MMT Stacked Ensemble Diagnoses of the Base Diagnosis Models.*

Features Used to Build the Models	**Base Models**	**Diagnoses Prediction Accuracy**			**False (Incorrect) Diagnoses Alarm Rate**		
		Base Model (%)	**MMT Model (%)**	**MMT Model Improvement over Base Model (%)**	**Base Model (%)**	**MMT Model (%)**	**MMT Model Improvement over Base Model (%)**
Consistency	DT	91.63	97.47	6.37	2.90	1.48	48.97
	NB	90.05		8.24	3.51		57.83
	KNN	86.41		12.80	4.78		69.04
Correlation	DT	90.84	95.58	5.22	2.93	2.24	23.55
	NB	87.52		9.21	3.99		43.86
	KNN	84.20		13.52	6.00		62.67
Information Gain	DT	93.36	99.05	6.09	2.66	0.37	86.09
	NB	91.31		8.48	3.09		88.03
	KNN	87.99		12.57	4.51		91.80
Whole Features	DT	81.83	90.68	10.82	5.82	3.84	34.02
	NB	78.67		15.27	7.19		46.59
	KNN	76.62		18.35	7.58		49.34

6. Conclusion

This research applied a stacked ensemble of three base models; KNN, C4.5 Decision Tree, and Naive Bayes, with Multiple Model Trees Meta classifier to improve the diagnosis of Lower Respiratory Tract Infection in Peadiatric. The results from this research establish the improvement of predictive model performances with feature selection techniques. The three reduced feature models' diagnosis accuracies were better than the accuracy recorded by the whole feature set model. Information gain models recorded the best accuracy, while correlation models recorded the least accuracy. Information gain is therefore endorsed to be used for features selection of diseases dataset. C4.5 Decision Tree-based models perform better than other base models across the three selected features

The MMT Meta model of the three base models of information gain selected attributes performs better than the MMT Meta models of the Consistency and the Correlation selected feature attributes. MMT Meta model with Consistency selected attributes performed better than Correlation selected attribute Meta model, the stacked ensemble of models built with the whole feature attributes recorded the least diagnoses accuracy. The Stacked Ensemble of diagnoses of models built with the Information Gain reduced feature attributes is therefore recommended to be used in health care delivery centers, especially in the rural areas where there are a shortage of medical doctors and qualified health personnel, for quick diagnosis and treatment of LRTI among pediatric patients to improve on health care delivery and save lives.

Conflict of Interest

The authors declare that they have no competing interests.

Acknowledgements

The authors acknowledge the following institutions and individuals for their support and contributions towards this research work's success:

a) Members of the Ethical Committee, and Staff of Record Department, Federal Medical Centre, Owo Ondo State, Nigeria.
b) Dr. O. Bello Head of Paediatrics Department, State Specialist Teaching Hospital, Akure, Nigeria.
c) Dr. Ajomale Abimbola, Medical Director, Federal Polytechnic, Ile Oluji, Nigeria.

References

[1] P. V. Dasaraju, C. Liu. "Infections of the Respiratory System". In: S. Baron, editor. Medical Microbiology. 4th edition. Galveston (TX): University of Texas Medical Branch at Galveston; 1996. Chapter 93. Available from: https://www.ncbi.nlm.nih.gov/books/NBK8142/ (Accessed: 26th July 2020).

[2] J. E. Crowe. "Viral Pneumonia. Kendig's Disorders of the Respiratory Tract in Children, 433–440. 2006. https://doi.org/10.1016/B978-0-7216-3695-5.50030-4.

[3] G. Worrall. "Acute bronchitis." Canadian family physician Medecin de Famille canadien, 54 (2), 238–239. 2008.

[4] K. Øymar, H. O. Skjerven, I. B. Mikalsen. "Acute bronchiolitis in infants, a review." Scandinavian journal of trauma, resuscitation, and emergency medicine, 22 (23). 2014. https://doi.org/10.1186/1757-7241-22-23

[5] S. A. Madhi, K. P. Klugman. "Acute Respiratory Infections." In: D. T. Jamison, R. G. Feachem, M. W. Makgoba, editors. Disease and Mortality in Sub-Saharan Africa. 2nd edition. Washington (DC): 2006. Chapter 11. Available from: https://www.ncbi.nlm.nih.gov/books/NBK2283/

[6] C. O. Oyejide, K. Osinusi, "Acute Respiratory Tract Infection in Children in Idikan Community, Ibadan, Nigeria: Severity, Risk Factors, and Frequency of Occurrence, Reviews of Infectious Diseases," Volume 12, Issue Supplement_8, Pages S1042–SI046, https://doi.org/10.1093/clinids/12.Supplement_8.S1042, 1990.

[7] R. Loddenkemper, G. J. Gibson, Y. Sibille. "Respiratory health and disease in Europe: the new European Lung White Book. European Respiratory"; European Respiratory Journal. 42: 559-563; DOI**:** 10.1183/09031936.00105513, 2013.

[8] A. D. Achary, K. S. Prasanna and S. Nail "Acute Respiratory Infections in Children: A Community Based Longitudinal Study in South India." Indian Journal of Public Health 47 (1): 1-13. 2003.

[9] T. Wardlaw, D. You, H. Newby, D. Anthony and M. Chopra "Child Survival: a Message of Hope but a call for Renewed Commitment in UNICEF report." Reprod Health".10 – 64. https://doi.org/10.1186/1742-4755-10-64. 2013.

[10] Yoo, Illhoi, P. Alafaireet, M. Marinov, K. Pena-Hernandez, R. Gopidi, J. Chang, L. Hua. Data mining in healthcare and biomedicine: a survey of the literature. Journal of medical systems 36 (4) 2431-2448. 2012.

[11] S. Apoorva, R. Pallavi, P. Kajal, S. R. Rai. "Health Analytics Using Machine Learning: A Survey." International Journal of Innovative Research in Computer and Communication Engineering Vol. 5, Issue 4, DOI: 10.15680/IJIRCCE.2017.05040116650,2017.

[12] H. Ameri, S. Alizadeh, E. Akhond Application of data mining techniques in clinical decision making: A literature review and classification. In: Akhond Zadeh Noughabi E, Raahemi B, Albadvi A, Far BH. Handbook of research on data science for effective healthcare practice and administration. IGI Global; 257-295. 2017. DOI: 10.4018/978-1-5225-2515-8.ch012.

[13] J, Ban, M. Gagliano, J. Pham, B. Tang, H. Kashif, "Applications of Machine Learning in Medical Diagnosis." Available from: https://www.researchgate.net/publication/321151498_Applications_of_Machine_Learning_in_Medical_Diagnosis, 2017. [Accessed Dec 05, 2019].

[14] Xu, J., Xue, K., & Zhang, K. "Current Status and Future Trends of Clinical Diagnoses via Image-Based Deep Learning." Theranostics, 9 (25), 7556–7565. https://doi.org/10.7150/thno.38065, 2019

[15] K. UmaMaheswari, A. Valarmathi, J. Jasmine., "Effective Diagnosis of Heart Disease through Stacking Approach. Advances in Natural and Applied Sciences". 11 (9); Pages: 323-328. 2017.

[16] A. Oguntimilehin, O. Adetunmbi, I. Osho "Towards Achieving Optimal Performance using Stacked Generalization Algorithm: A Case Study of Clinical Diagnosis of Malaria Fever" The International Arab Journal of Information Technology, Vol. 16, No. 6. 2019.

[17] O. C. Olayemi, O. S. Adewale, O. O Olasehinde, B. A. Ojokoh, A. O. Adetunmbi. "Application of Machine Learning to the Diagnosis of Lower Respiratory Tract Infections in Paediatric Patients." i-manager's Journal on Pattern Recognition, 5 (2), 21-29, https://doi.org/10.26634/jpr.5.2.15538, 2018

[18] M. Jan, A. A. Awan, M. S. Khalid, S. Nisar, "Ensemble Approach for Developing a Smart Heart Disease Prediction System using Classification Algorithms." Research Reports in Clinical Cardiology, 9: 33-45, 2018.

[19] R. J. McDonough, "Utilizing Data Mining Techniques and Ensemble Learning to Predict Development of Surgical Site Infections in Gynecologic Cancer Patients" Graduate Dissertations and Theses. 33. https://orb.binghamton.edu/dissertation_and_theses/33, 2018.

[20] J. R. Quinlan," C4.5: Programs for Machine Learning." San Francisco: Morgan Kaufmann. Publishers, Inc., https://dl.acm.org/doi/book/10.5555/152181, 2003

[21] O. O. Olasehinde, O. C. Olayemi, B. K. Alese. " Multiple Model Tree Meta Algorithms Improvement of Network Intrusion Detection Predictions Accuracy" International Journal for Information Security Research (IJISR), Volume 9, Issue 3, https://infonomics-society.org/wp-content/uploads/Multiple-Model-Tree-Meta-Algorithms-Improvement-of-Network-Intrusion-Detection.pdf, 2019.

[22] A. Kaveh, S. M. Amze-Ziabari, T. Bakhshpoori. M5' Algorithm for Shear Strength Prediction of HSC Slender Beams without Web Reinforcement. IJMO 7 (1), 2017. DOI: 10.1617/s11527-015-0752-x.

[23] H. Duggal and P. Singh, "Comparative Study of the Performance of M5-Rules Algorithm with Different Algorithms," Journal of Software Engineering and Applications, 5 (4) 270-276. doi: 10.4236/jsea.2012.54032.2012.

[24] P. Yildirim, "Filter-Based Feature Selection Methods for Prediction of Risks in Hepatitis Disease." International Journal of Machine Learning and Computing 5 (4): 258 – 263, 2015.

[25] O. O. Olasehinde, B. K. Alese, A. O Adetunmbi, Performance Evaluation of Bayesian Classifier on Filter-Based Feature Selection Techniques, International Journal of Computer Science and Telecommunications 9 (7) (2018) 24-30.

[26] Z. Shichao, L. Xuelong, Z. Ming, Z. Xjaofeng, C. Debo, "Learning K for KNN Classification," ACM Transactions on Intelligent Systems and Technology, https://doi.org/10.1145/2990508, January 2017.

[27] C. Gaurav "All about Naive Bayes "Towards Data Science," A Medium publication sharing concepts, ideas, and codes https://towardsdatascience.com/all-about-naive-bayes-8e13cef044cf, 2018. (Accessed on 6th June 2020).

[28] B. N. Lakshmi, T. S. Indumathi, R. N. Ravi, "A study on C4.5 Decision Tree Classification Algorithm for Risk Predictions during Pregnancy", International Conference on Emerging Trends in Engineering, Science and Technology (ICETEST - 2015), Procedia Technology. (24) 1542-1549, 2016.

[29] E. Frank, Y. Wang, S. Inglis, G. Holmes, I. H. Witten. Using Model Trees for Classification. Machine Learning, (32): 63-76, 1998.