IJCSN - International Journal of Computer Science and Network, Volume 8, Issue 5, October 2019
ISSN (Online) : 2277-5420
www.IJCSN.org
Impact Factor: 1.5

Stacked Ensemble Approach to the Development of Lower Respiratory Tract Infection Diagnoses System

[1] Olayemi OLASEHINDE; [2] Olufunke OLAYEMI

[1] *Department of Computer Science, Federal Polytechnic*
Ile Oluji, Ondo State, Nigeria

[2] *Department of Computer Science, Joseph Ayo Babalola University*
Ikeji-Arakeji, Osun State, Nigeria

Abstract - Lower Respiratory Tract Infections (LRTIs) is the second leading cause of death among paediatric patients in Nigeria. It is also ranked as the third leading cause of death in the United states. Machine learning techniques has been widely applied to diagnose diseases and infections all over the world. Stack Ensemble has been used to improve machine learning diagnosis by combining diagnosis of several machine learning models. In this work, three machine learning techniques (Naive Bayes', K-Nearest Neighbour (KNN) and Decision Tree Algorithm) are used to build base diagnoses models of each of the three reduced selected features of the LRTIs dataset; consistency, correlation and information Gain feature selection techniques and the whole feature attributes of the dataset, the diagnoses of each of the base models built from each of the reduced feature and whole feature attributes are combined with Multiple Model Trees (MMT) Stacked Ensemble. The Diagnosis models of Information Gain reduced features set recorded the highest diagnosis accuracy and lowest wrong diagnosis rate, followed by consistency reduced features set, while correlation reduced features set recorded the least diagnosis performances ahead of the whole features set models. The MMT model recorded highest diagnoses accuracy improvement with the KNN models; 12.80% for consistency feature model, 13.52% for correlation feature model, 12.37% for information Gain feature model, and 18.35% for whole feature model, MMT model recorded lowest improvement with Decision Tree models; 6.37% for consistency feature model, 5.22% for correlation feature model, 6.09% for information Gain feature model, and 10.82% for whole feature model. In terms of False Diagnoses Alarm Rate, KNN also recorded highest improvements; 69.04% for consistency feature model, 62.67% for correlation feature model, 91.80% for information Gain feature model, and 49.34% for whole feature model while it recorded lowest improvement with Decision Tree models;48.97% for consistency feature model, 23.55% for correlation feature model, 86.09% for information Gain feature model, and 34.02% for whole feature model.

Keywords - Lower Respiratory Tract Infection (LRTI), Diagnosis, Machine Leaning, Stacked Ensemble, Improvement

1. Introduction

Lower Respiratory tract infections (LRTIs) are infections along the respiratory tract of human body [1], caused by bacteria, fungi and virus. Clinicians describe LRTIs with the use of a wide range of disease definitions, such as acute bronchitis, bronchiolitis and pneumonia, depending on the symptoms and signs that can be observed. Viral infection are the main causes of mild and moderate pneumonia (especially in the first year of life) while bacterial infections are the leading cause of severe pneumonia [2]. Respiratory diseases among children are major cause of mortality and morbidity worldwide. It is the second leading killer of young children around the world due to the attitude of care givers, causing approximately20% of all child's deaths. [3] and constituted the second leading cause of death in all age bracket in Nigeria, According to [4], LRTIs is the third largest cause of death in the United States. On the average, research have shown that a child can be infected with LRTIs up to six times in a year, accounting for about 30 % - 50% of the total paediatrics outpatient visits [5]. Pneumonia is part of the primary cause of morbidity and mortality among children under the ages of 5 years of age in the developing Countries [6]

The risk of infections are on the increase, there are lots of risk factors which are responsible for respiratory infections. Children cannot avoid having contacts with viruses and bacteria which are the first risk factors in increasing their chances of developing respiratory infections. The resistant systems of children are more prone to being affected by the infection. Children regular contact with other children who could be infection carriers

always puts them at risk. Also, Most times, children often don't 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 the infection ,early detection and treatment of this infection will reduce the mortality rate caused by it. Application of data mining techniques will aid early diagnoses of this infection in children. Data mining are automatic techniques for discovering hidden and potentially useful patterns from a dataset and generate new information from it that can be a useful tool in the health sector and healthcare[7]. Predicting the health status of a patient is an interesting application areas of data mining. Classification is a data mining technique used to predict group membership for patient risk factors. In the case of Lower respiratory tract infection (LRTI), data mining is used to build diagnosis model that predicts health status of a patient as either infected or not infected with LRTI. Building of the LRTI predictive model involves the training of Machine Learning Algorithms with LRTI training dataset and its evaluation with LRTI test dataset. Stacked Ensemble are techniques for improving the results of individual evaluated predictive models, Ensemble Learning are machine learning techniques that improve the diagnoses of predictive models by combining the diagnoses of such several models into a single stacked model with better predictive performance than the best performance recoded by any of the combined predictive models[8].This paper uses Multiple Model Tree meta learners algorithm to combine and improve the diagnosis of three LRTI models built from three data mining techniques; k-nearest Neighbour Classifier (KNN), Naïve Bayes' (NB) and Decision tree algorithms.

2. Literature Review

Olayemi et. al., presented a paper titled "Application of Machine Learning to the Diagnosis of Lower Respiratory Tract Infections (LTRIs) in Paediatric Patients"[9],This paper employed the use LRTIs clinical dataset to train two machine learning techniques; Naïve Bayes and K- nearest neighbor and built models used to diagnose the presence of LRTI in infants. The performance of the models was evaluated based on accuracy, sensitivity, specificity and precision. The result of Naïve Bayes and k-nearest neighbor with all features (18) used shows 94.25% and 94.43% respectively. Naïve Bayes with information- based feature selection method shows accuracy of 99.60% while k-nearest neighbor shows 94.35% with 10 features. Also, Naïve Bayes with Correlation-based feature selection method shows accuracy of 95.40% while k-nearest neighbor shows 95.40% too with just six (6) features. The comparative results show that Naïve Bayes with information based feature selection method performs stronger and better than others. [10] presented a paper titled '' the development of a predictive model for paediatric 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 in carrying out the predictive task with minimum 92% accuracy. [11] reported that stacking with Meta Decision Tree (MDTs) clearly outperforms voting and stacking with decision trees, as well as boosting and bagging of decision trees. On the other hand, MDTs performed only slightly better than SCANN and selecting the best classifier with cross validation (SelectBest) .[12], investigated classification via regression, and reported that classification via Multi Response Model Trees (MMT) performs extremely better than multi response linear regression (MLR) and better than C5.0 (a successor of C4.5, especially in domains with continuous attributes. This indicates that multiple model trees (MMT) are a very suitable choice for learning at the meta-level. [13] investigated the possibility of using ensemble algorithms to improve the performance of network intrusion detection systems. Three machine learning algorithms; K Nearest Neighbour, 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 Multiple Model Tree (MMT)meta learner algorithm individually and is collectively evaluated using a ten-fold cross validation to build the stacked ensemble model used for classifications of the network traffics into any of the nine network attacks or normal. The results from this research show that MMT stacked model of the three base learner predictions gives a higher multi-class classification accuracy than the best accuracy recorded by any of three base models prediction for each of the network connection categories. 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 base models and the MMT stacked model.

3. Lower Respiratory Tract Infections (LTRIs) Dataset

A total of two thousand, one hundred and ten (2110) instances were collected from November,2015 to October,2016, Purposive data collection technique was used and a direct access to the medical records of patients clinically diagnosed to have respiratory infections was granted by authorities of the health institution. one

thousand, eight hundred and fifty four (1854) instances out of the 2110 instances were clinically diagnosed to have respiratory infections, which consist of one thousand, five hundred and twenty (81.98%) records of Pneumonia patients, one hundred and eleven (5.99%) records of Bronchitis patients and two hundred and twenty three (12.03%) records of Bronchitis patients, all these patient records were stored as infected with LTRI. The remaining two hundred and fifty six (256) instances out of the 2110 collected instances were clinically diagnosed to have other infections such as HIV, Tuberculoses, Typhoid Fever, Malaria Fever etc. These categories were stored as Not Infected with LTRI. The data were collected in a raw form and later stored in the MS-Excel format for further pre-processing exercises. The total datasets were splited into two, seventy (70% - 1477 instances) for training and thirty (30% - 633 instances) for testing set. the training. The training dataset consists of 1314 instances Infected with LRTI and 163 Instances not Infected with LRTI, the testing dataset consists of 540 instances Infected with LRTI and 93 instances not Infected with LRTI as shown in Table 1.

Table 1: Distribution LRTI Dataset Splits

	Training Dataset (70%	*Testing Dataset (30%)*	*Total*
LRTI Infected	*1314*	*540*	*1854*
Non LRTI Infected	*163*	*93*	*256*
Total	*1477*	*633*	*2110*

Pre-processing were carried out on the data from the raw form to Microsoft Excel format,All the attributes were discretised with variable field length as shown in Table 2. There are eighteen (18) conditional attributes and one (1) class identity.Each conditional attributes was assigned a value of Yes or No, depending on the patients symptoms . Yes (1) means the symptom is present in the patient while No (0) means the symptom is not present in the patients body. The combination of the available attributes/features (Symptoms) gives room for classification (diagnosis) of each record. This decision attributes denote the results of the diagnosis carried out. Based on the conditional attributes, each record was assigned a class by the medical experts. There are also two classes in the respiratory infection case, Yes (1) means the LRTI infection is present in the patients while No (0) means the LRTI infection is not present in the patient's body.

A detailed description of each of the attributes in the dataset is as follows:

a) **Sex:** is a description of the gender of the patient 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:** is a description of the weight of patients 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) **Breast feeding:** is the description of sucking babies. Breastfeeding until the age of 4months and partially thereafter was associated with a significant reduction of respiratory and gastrointestinal morbidity in infants. The value can either be yes or no
e) **Parental smoking:** this is a description of parents that smoke.
f) **Cyanosis** is the appearance of a blue or purple coloration of the skin or mucous membranes due to the tissues near the skin surface having low oxygen saturation.
g) **Respiratory Rate**: this rate is usually measured when a person is at rest and simply Involves counting the number of breaths for one minute by counting how many times the chest rises. Respiration rates may increase with fever, illness, and with other medical conditions.
h) **Cough** is a very common illness in for children. It is a reflex that helps clear the airway for secretions, protects the airway from foreign body aspiration, and can be a sign of manifesting symptom of a disease. The most common cause of cough is cold.
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) (Medlineplus Medical Encyclopedia)
j) **Indoor air pollution exposure**: This is an exposure to indoor air pollution (use of wood and biomass fuels) for cooking.
k) I**mmunization** is the vaccination against some diseases in children. Lack of or incomplete Immunization for children can result in serious diseases in children

l) **Crowding.** This occurs when there are more than 7 persons per bedroom (8 by 10 room). Also when there are more than 4 persons sharing a child's bedroom according to (WHO, 2012). The room is said to be overcrowded.
m) **Fever:** If a child has high temperature or low body temperature, the child is likely to have 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 as the result of exercise or nasal congestion. It may be a sign of a more serious heart infection.
p) **Herbal mixture**: Herbal mixtures are 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 together and given to a sick child or person like a medicine.
q) **Educational Status**: This is the level of educational background of the parents or caregiver
r) **Daycare**: This is daytime care for people who cannot be fully independent, such as Children or elderly people. Children would be looked after in day care 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.

4. Methodology

Figure 1 shows the architecture of the proposed Stacked Ensemble of Respiratory Tract Infection diagnosis In Paediatric using Multiple Model Trees Meta Algorithm, it consists of three different stages; the dataset discretisation stage handles the discretisation of the LRTIs dataset. The second stage involves the building of the base predictive models, the three-base algorithm; K Nearest Neighbour, Decision tree and Naïve Bayes' were trained with the training set and evaluated using test data to generate the base diagnoses. The last stage involves the training and building of the MMT ensemble model, the diagnoses of the three base models were used to train and evaluates by Multiple Model Tree (MMT) meta algorithm via ten folds cross validation to build the MMT stacked ensemble model used for the meta level classifications of the patients as either LRTIs infected or not Infected. The system was implemented with Python language on a Corel i3, 64bits, 2.4 GHz processor, 16MB Cache, 512GB SDD, Ms. Windows 7 operating system.

4.1 Feature Selection

Feature Selection (FS) is important in machine learning tasks because it can significantly improve the performance y eliminating redundant and irrelevant features at the same time speeding up the learning task [14][15]. Identification of the variables relevant to LRTIs prediction is likely to improve the performance of the supervised machine learning algorithm's performance and also reduce the models complexity [16]. Three Filter-based FSmethods; Consistency, Correlation and Information Gain were used in this study to determine the relevant features among the risk factors present in the collected LRTI dataset.

Table 2 Attrbutes, Description of Variables and Meaurements of LRTIs

S/N	*ABBREVIATION*	*ATTRIBUTES*	*ATTRIBUTES TYPE*	*Description*
1	*SEX*	*SEX*	*Discrete*	*Male/female*
2	*AGE*	*AGE*	*Discrete*	*Yes = 0-5 years, No= 5 – adult*
3	*BRFD*	*BREAST FEEDING 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*	*Normal <=36 yes , High >37.5 no*
8	*RR*	*RESPIRATORY RATE*	*Discrete*	*Normal =35 yes, High >40 cy/m no*
9	*HRRR*	*Heart RATE*	*Discrete*	*Normal = 1* *High = 0*
10	*POLL*	*INDOOR AIR POLLUTION*	*Discrete*	*1=Yes, 0= No*

11	*IMMU*	*INCOMPLETE IMMUNIZATION*	*Discrete*	*1 = yes Complete immunization* *0 = no not complete 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*

Consistency Based FS

Consistency Based FS generate all possible features subset of the LRTI dataset, for each of the generated subset, compute the inconsistency count INC_i of all pattern p_i of the subset S, then calculate the inconsistency rate INCR of subset S using Equation 1, select and return subset with the lowest inconsistency rate INCR

$$(C, D \mid F) = \frac{\sum_{i=1}^{n} \frac{max_i^2}{m_{ir}}}{n} \quad (1)$$

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

Correlation Based FS

Correlation based FS generates all possible attributes (features) subset S of the LRTI dataset , then calculate $Merit_s$ for each of the subset S, contain k features using Equation 2, subset S with the highest $Merit_s$ value is selected and returned

$$M_s = \frac{k\,\overline{r_{cf}}}{\sqrt{k+k(k-1)\overline{r_{ff}}}} \quad (2)$$

where; $\overline{r_{cf}}$ = average feature-class correlation
$\overline{r_{ff}}$ = average feature-feature correlation

Information Gain Based FS

Information Gain features Selector computes the Information Gain (IG) for each feature of the LRTI dataset, ranks each of the feature based on their IG value in descending order and validates set of the ranked attributes in terms of classification accuracy on the training dataset, select and return set of validated attributes with highest classification accuracy

$$IG = H(Y) - H(Y|X) = H(X) - H(X|Y) = H(Y) + H(X) - H(X|Y) \quad (3)$$

Where H(Y) is Entropy of Y denoted in Equation 4 and H(Y | X) is Entropy of Y given X denoted in Equation 5

$$H(Y) = -\sum_{y \in Y} p(y) \log_2(p(y)) \quad (4)$$

$$H(Y|X) = -\sum_{x \in X} p(x) \sum_{y \in Y} p(y|x) \log_2(p(y|x)) \quad (5)$$

Where n: is number of instances in the LRTI dataset, k: is the categories of class label in the LRTI dataset, $P(y_i)$: is the probability of occurrence of attack categories value of instance i

4.2 Base Models

Three machine learning algorithms; K Nearest Neighbor, Naive Bayes and Decision Tree were adapted to build the base models

K Nearest Neighbor

K-Nearest Neighbor is based on Euclidean distance between the training set and the testing set, it classifies an unlabeled instance (tuple) in the LRTI test dataset by assigning it to the class of the most similar labeled instances in the LRTIs training dataset. It uses distance function in Equation 6 to compute the distance between the given instance and all the instances in the LRTI training dataset, rank the computed distances of all the training instances with the given instance in ascending order, then uses simple majority vote to determine instance that has highest number of class id (Infected or not infected) among the topmost ranked patients, then Classifies the given patients class id as the majority class id among the topmost ranked patients

$$d(p_i, q_t) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \quad (6)$$

From equation (6), a given instance will be classified as the class label category having majority class label value among closest instance to the given instance

Naive Bayes

Naïve Bayes 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 class id of a patient given its risk factors; it calculates the probability for each attribute, conditional on the each class id, use product rule to obtain a joint conditional probability for all the patient risk factors, then use Bayes rule to derive conditional probability for each class label and predict the class label of the given patient as the class label with the highest probability. Theprobability that a class label y_j will be assigned to a given unlabeled instance X of the LRTI dataset is given in Equation 7.

$$p(y_j \mid x_1, \ldots, x_{43}) = \frac{p(y_j)p(x_i \mid y_j)}{p(x_i)} \; (\forall_j = 0,1.,9) \qquad (7)$$

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

$$y = \arg\max_{y} py_j \prod_{j=0}^{9} p(y_j)p(x_1, x_2, \ldots x_{43} \mid y_j) \qquad (8)$$

Decision Tree

Decision Tree (DT), (C4.5) is a classification model consisting nodes that are attributes of LRTIs dataset and arcs which are attribute values connection to other nodes all the way to the leaves which are the class id (class label), it builds a classification tree, which will be used to predict the class id of a new patients; DT calculates the Gain Ratio of all the attributes (X) of the training dataset as shown in Equation 9, the Split value is given in Equation 10. The Attribute with the highest Gain Ratio is use to divide the dataset attribute into two subsets, the attribute with the highest gain ration of the two subsets is further used to divide each of the subset 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 satisfies her risk factors attribute values, The information Gain of attribute X is given in Equation 3

$$\text{Gain Ratio} = \frac{Information\ Gain(X)}{Split\ information(X)} \qquad (9)$$

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

Where n is the number of values in attribute X and $|\mathbf{x}|$ is the value of the attribute X

4.3 Stacking with Multiple Model Trees. (MMT)

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 class id, will be used to train machine learning algorithms $L_1, . , L_k$ to create base classifiers $C_1, C_2, \ldots, C_k$, where $C_i = L_i(S)$. The diagnosis of the base models will be of the form $\hat{y}_1$, $\hat{y}_{2,..}$ $\hat{y}_{k.}$, where k is the number of base classifier. In the second phase, the diagnosis results of the base models will be used to train the meta learner algorithm using leave-one-out validation technique, the meta learner L_i^k will be applied to the entire base classifiers diagnosis dataset $\hat{S}$, leaving one fold $\hat{s}_i$ out for testing, $C_i^k = L_i^k(\hat{S} - \hat{s}_i)$. The learned meta classifiers will generate diagnosis of the form $C_i^k(\hat{s}_i)$. Multiple model trees is a meta learner adaptation of decision tree with linear regression functions at the leaves. it can be adapted to diagnose and applied to classification problems by employing a standard method of transforming a classification problem in to a problem of function approximation, 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.

Figure 2, show the algorithm of MMT while Figure 3 shows the pictorial representation of the operations of MMT stacking, from the Figure 3, MMT first generated a derived dataset based on the distinct class label from the main LRTIs dataset, trees generated from the derived dataset were induced with M5' linear regression algorithm to create a classification function for class id, the values of the risk factors of a patient to be diagnosed were plugged into each of the linear regression function, the function with the highest value is returned and the patients will be diagnosed as class id of the function with highest value

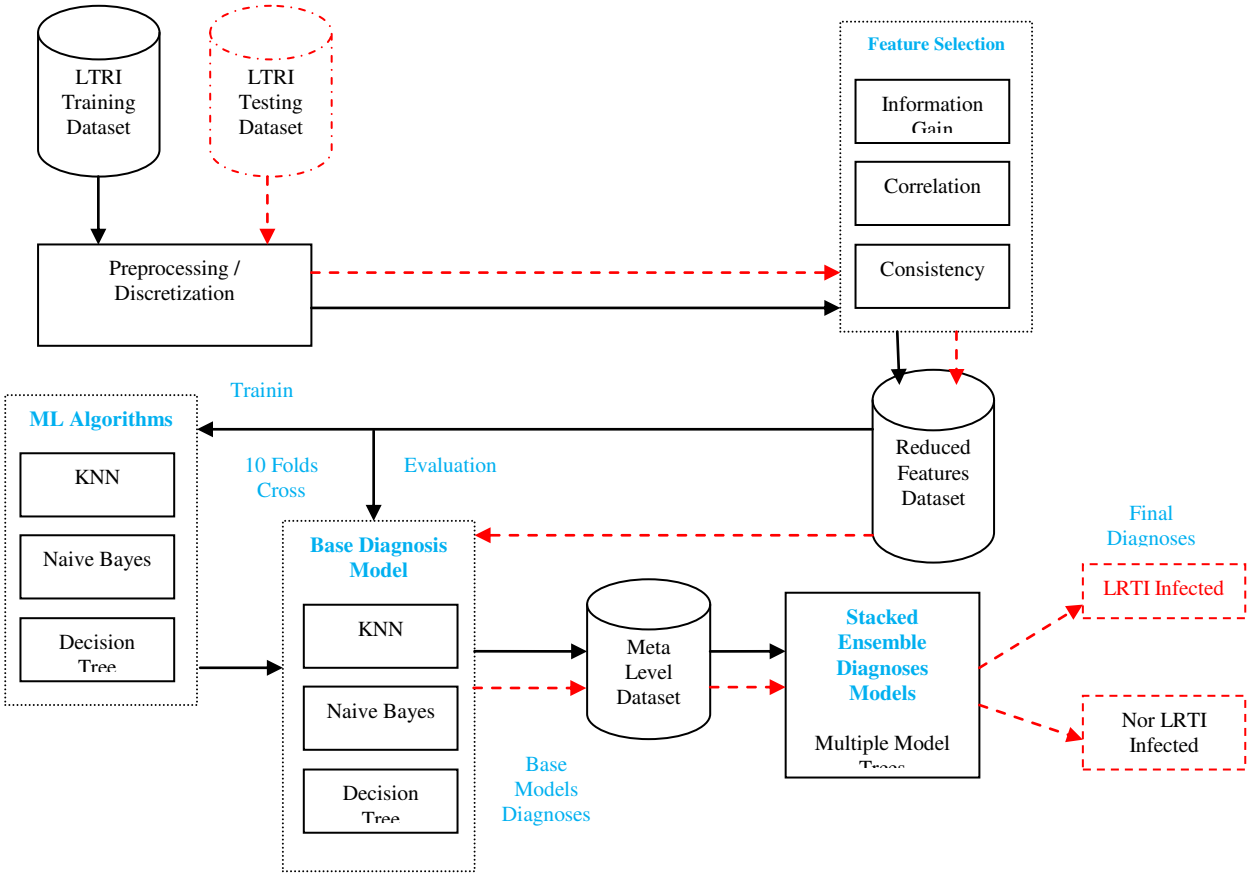


Figure 1: The Architecture of the Stacked Ensemble of Lower Respiratory Tract Infection Diagnosis System

input: *Data set D (LRTIs) ={ (x_1,y_1), (x_2,y_2),.... , (x_m,y_m)}*
new instance to be classifier(x_1, x_2,.....x_n)

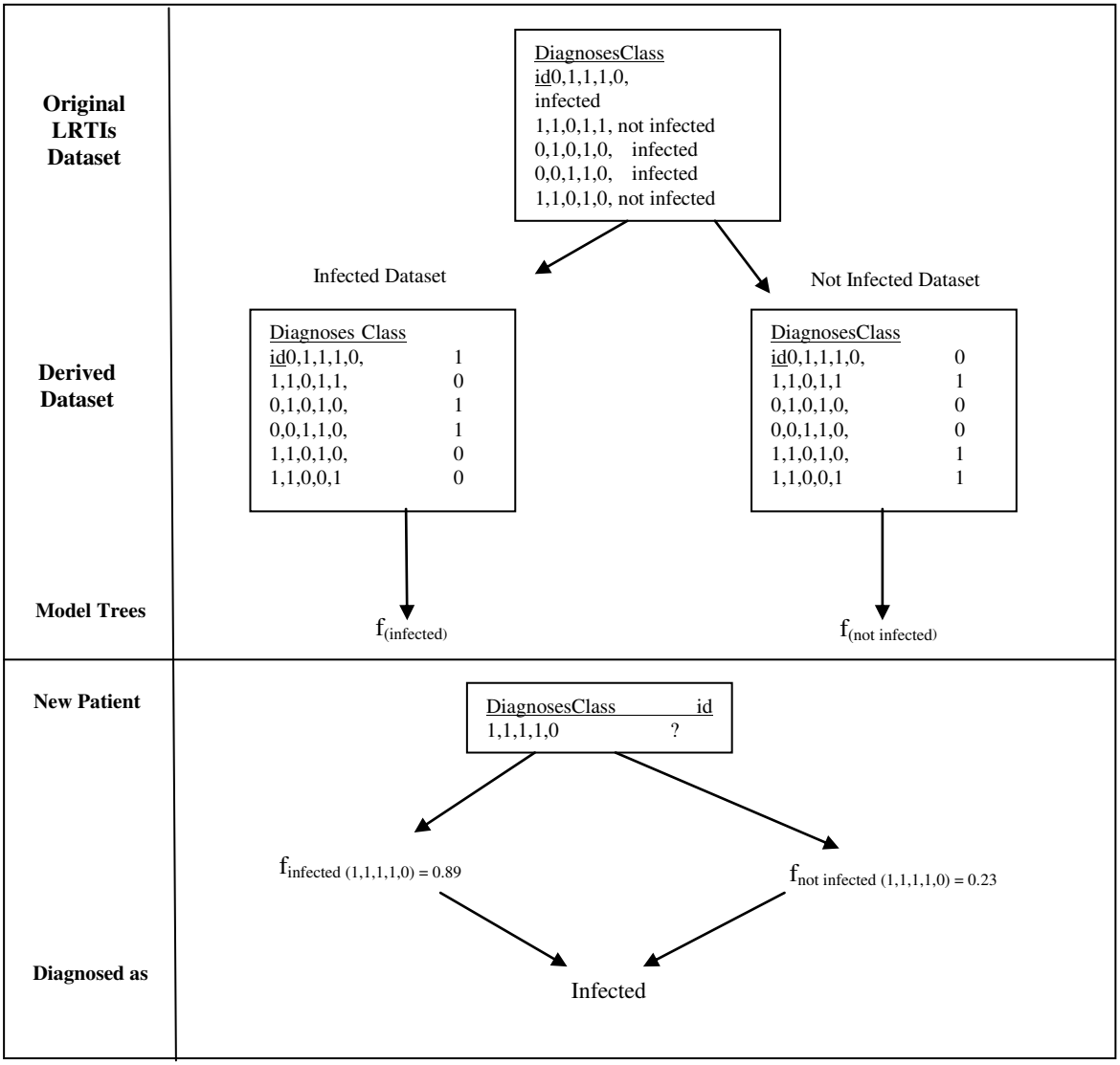
Process: *for i = 1...,n; // n is distinct number of class id*
$D_i' = D$ I] {((x_{i1}, x_{i2},....x_{in}), y_i)} // generate a derived dataset for each of the
// distinct class id)
end;

for i = 1...,n;
$f_i = m5'(D_i')$ // create a linear function for each of the distinct class id
// by inducing each of the derived dataset with the m5' linear
// algorithm
end;

for k = 1...,n;
Value_f_k = f_k ((x_1, x_2,.....x_n)
end;

Output: $= \arg \underset{value\_f_k}{maximum} = f_k(x_1, x_2, \ldots\ldots, x_n)$
end;

Figure 2: Multiple Model Trees algorithm



5. Results and Discussions

Table 3 shows 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. The results of the data collected showed that majority of the patients were male with a proportion of 66.39% owing for a 2 to 1 ratio for male to female records, this shows that male children are more vulnerable to illness than female child, the age distribution showed that majority of the patients were below a year with a proportion of 84.84% of patients. The results also showed that majority of the patients have difficulty in breathing which is 54.78% and majority of the patients do not have cyanosis with a proportion of 83.70%. It further showed that majority of the patients have normal body weight with a proportion of 82.99% and the remaining with below normal weight while a majority of the patients treated had abnormal temperatures with a proportion of about 52.18%. The results also showed that majority of the patients had cough represented by a proportion of 89.19%. 80.38% had fever. Majority of the patients also had very abnormal respiratory rate (RR) represented by a proportion of 72.60% while a majority of the patients were also observed to have pollution and represented by a proportion of about 69.48%. The results also shows almost equal number of patients were observed for those with and without Immunization and whose parents were either smokers or educated. 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 day care centers. The results further showed that majority of the patients were administered Herbal mixture with a proportion of about 62.80% while majority 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 showed 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
Immunisation	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 shows the result of the feature selection techniques used in this study, consistency-based feature selection approach showed that there were twelve (12) relevant features, the information-based feature selection results showed that there were

ten (10) relevant features selected, the correlation-based feature selection algorithm was used to select the subset of features highly correlated with the target class (LRTI) but with lower correlation with other features. Using this strategy to identify relevant features showed that there were six (6) important 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 shows confusion matrix and the diagnoses accuracy and false diagnoses rate of the base and ensemble diagnosis models for each of the three reduced attributes and whole attributes models. Decision tree models recorded the highest diagnosis accuracy across all the features set among the base models; 91.63% with consistency reduced features set, 90.84% with correlation reduced feature set and 93.36% with information gain reduced set, closely followed by Naive Bayes models; 90.05% with consistency reduced features set, 87.52% with correlation reduced feature set and 91.31% with information gain reduced set, KNN models recorded the least diagnosis accuracy of 86.41% with consistency reduced features set, 84.20% with correlation reduced feature set and 87.99% with information gain reduced set among the base models. The MMT ensemble of each of the base models diagnosis, recorded highest diagnosis accuracy for each of the features set; 97.47% for consistency reduced set, 95.58% for correlation reduced feature set, 99.05% for the information gain reduced features set and 90.68% for the whole features set. Diagnosis models of Information Gain reduced features set recorded the highest diagnosis accuracy and lowest wrong diagnosis rate, followed by consistency reduced features set, while correlation reduced features set recorded the least diagnosis performances ahead of the whole features set models. Figures 5 and 6 show the performances of all the models in term of diagnosis accuracy and wrong diagnoses rate respectively.

Table 6 shows 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 highest diagnoses accuracy improvement with the KNN models; 12.80% for consistency feature model, 13.52% for correlation feature model, 12.37% for information Gain feature model, and 18.35% for whole feature model, MMT model recorded lowest improvement with Decision Tree models; 6.37% for consistency feature model, 5.22% for correlation feature model, 6.09% for information Gain feature model, and 10.82% for whole feature model. In terms of False Diagnoses Alarm Rate, KNN also recorded highest improvements; 69.04% for consistency feature model, 62.67% for correlation feature model, 91.80% for information Gain feature model, and 49.34% for whole feature model while it recorded lowest improvement with Decision Tree models;48.97% for consistency feature model, 23.55% for correlation feature model, 86.09% for information Gain feature model, and 34.02% for whole feature model, Figure 7 and Figure 8 shows the MMT Diagnoses Accuracy Improvement over Base models and False (Incorrect) Diagnoses Alarm Rate Improvement over Base models respectively, Figure 7 and Figure 8 shows the Graphical presentation of the performance of the MMT Stacked Ensemble over the base models in term of Diagnoses Accuracy and False (incorrect) Diagnoses Rate respectively

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

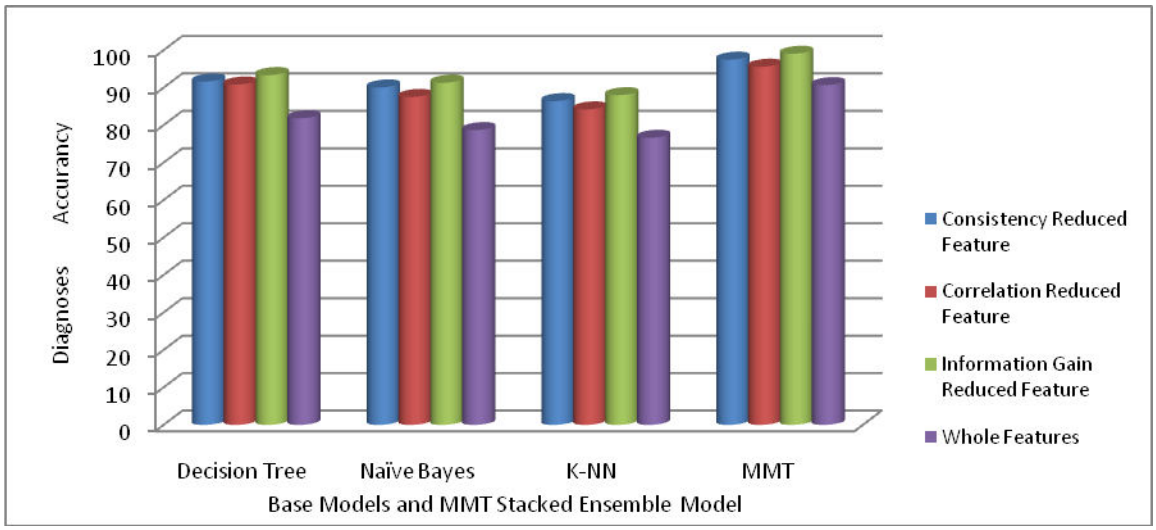


Figure 5: Diagnoses Accuracy Performance recorded by the Base and the Ensemble Models

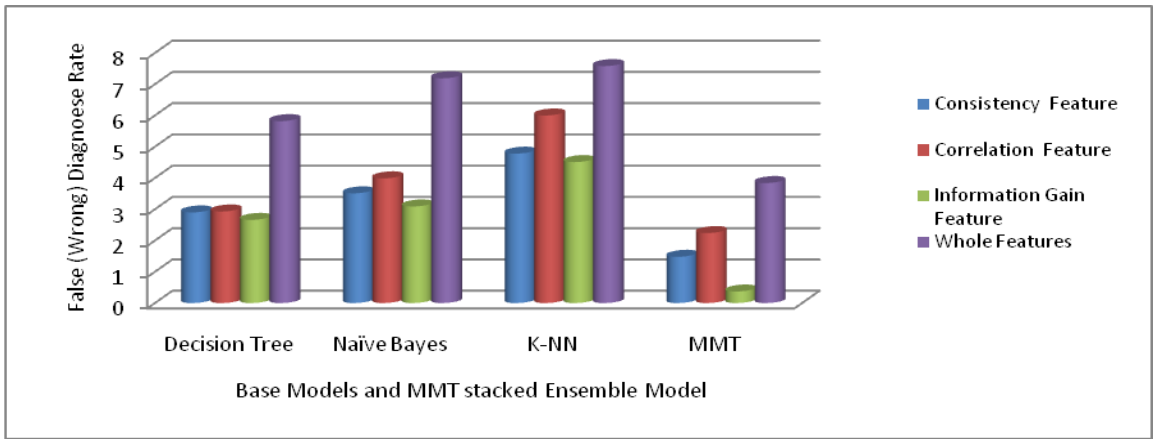


Figure 6: False (Wrong) Diagnoses Rate Performance Recorded by the Base and the Ensemble Models

Table 6: Diagnoses Accuracy and False (Wrong) Diagnoses Alarm Rate Improvement of MMT Stacked Ensemble Diagnoses Over the Base Models Diagnoses

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		6.37	2.90		48.97
	NB	90.05		8.24	3.51		57.83
	KNN	86.41	97.47	12.80	4.78	1.48	69.04
Correlation	DT	90.84		5.22	2.93		23.55
	NB	87.52		9.21	3.99		43.86
	KNN	84.20	95.58	13.52	6.00	2.24	62.67
Information Gain	DT	93.36		6.09	2.66		86.09
	NB	91.31		8.48	3.09		88.03
	KNN	87.99	99.05	12.57	4.51	0.37	91.80
Whole Features	DT	81.83		10.82	5.82		34.02
	NB	78.67		15.27	7.19		46.59
	KNN	76.62	90.68	18.35	7.58	3.84	49.34

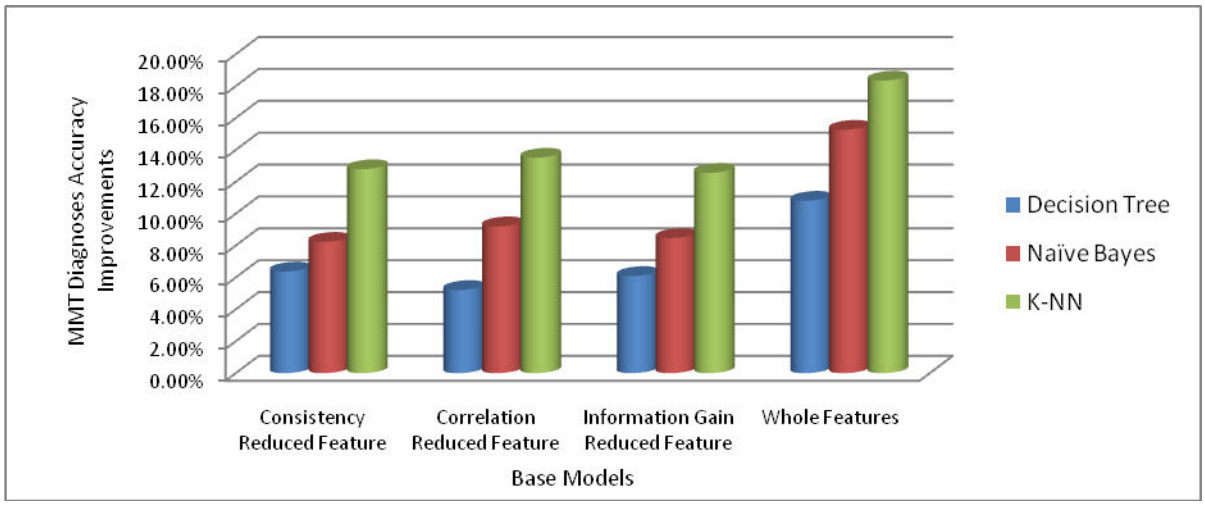


Figure 7: MMT Diagnoses Accuracy Improvement over Base models

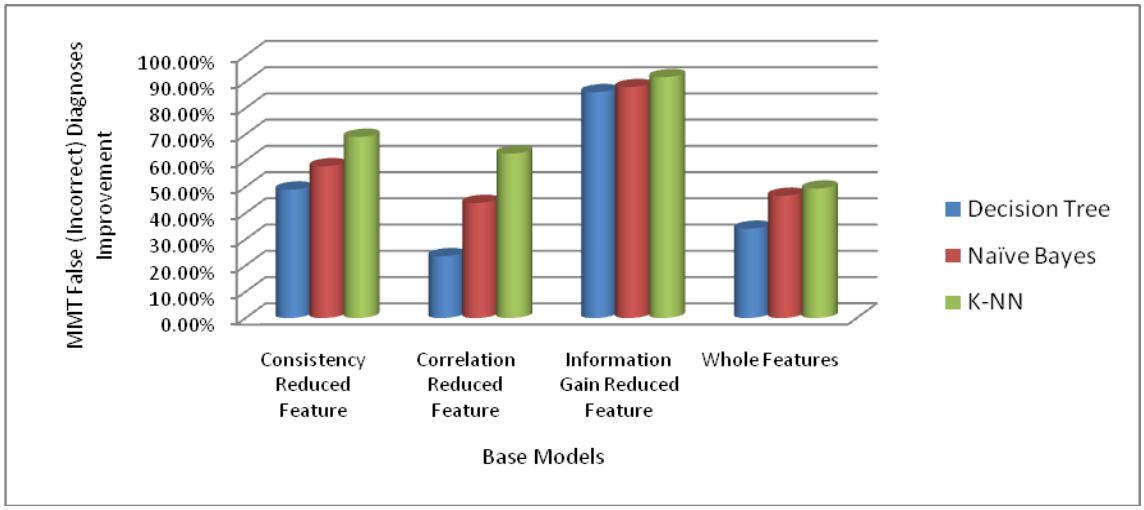


Figure 8: MMT False (Incorrect) Diagnoses Alarm Rate Improvement over Base models

6. Conclusion

In this paper, Multiple Model Tree Meta Algorithm has been used to develop a Stacked Ensemble Lower Respiratory Tract Infection Diagnoses System that combines the Diagnoses of three base models; KNN, Decision Tree and Naive Bayes. This work establishes a better diagnoses of LRTI with the Base models developed with Reduced Feature selected attributes of the LRTI Dataset than the model developed with its whole feature attributes, Base models developed with the Information Gain reduced features recorded highest diagnoses accuracy closely followed by Models developed from consistency reduced feature attributes, models of correlation reduced feature attributes recorded the least diagnoses accuracy, Information Gain feature selection techniques is therefore recommended for dimensional and complexity reduction of infection dataset. Decision Tree models recorded higher diagnoses accuracy than Naive Bayes and KNN base models. The MMT Stacked Ensemble of the diagnoese of the base models with the information gain reduced selected features recorded highest diagnoses accuracy of 99.05%, followed by the Stacked Ensemble of consistency models which recorded 97.47% diagnoses accuracy, the Stacked Ensemble of Correlation models recorded the diagnoses accuracy of 95.58% ahead of the Stacked Ensemble of models with the whole feature attributes of 90.68% 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 centres especially in the rural areas where there are shortage of medical doctors and qualified health personnel for quick

diagnoses and treatment of LRTI among paediatrics so as improve on health care delivery and save lives.

Acknowledgements

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

Ethical Standard
Funding: This research work is a self funded research undertaken by the authorsat the Department of Computer Science, School of Computing, Federal University of Technology, Akure, Ondo State, Nigeria

Conflict of Interest: The corresponding author states that there is no conflict of interest

References

[1] World Health Organization “Taking stock : Health workers shortage and the Response to AIDS”, www.who.int/healthsystems/.../TTR. 2008
[2] World Health Organization (2012).Pneumonia, fact sheet . Technical report, The Health Resources and Services Administration’s Maternal and Child Health Bureau (MCHB, 2011)
[3] Health Line Editorial Team a Fair Chance for Every Child (2017)
[4] Loddenkemper R, Gibson GJ, Sibille Y. Respiratory health and disease in Europe: the new EuropeanLung White Book. European Respiratory; 2013.
[5] A. D. Achary, K. S. Prasanna and S. Nail “Acute Respiratory Infections in Children: A Community Based Longitudinal Study in South India.” Indian J. Public Health 47(1):1 -13. January 2003
[6] T. Wardlaw, D. You, H. Newby, D. Anthony and M. Chopra “Child survival: a message of hope but a cal for renewed commitment inUNICEF report”. Reprod Health”. Pp.10 – 64 , May 2013
[7] R. E. Schapire, "The strength of weak learnability. Mach Learn " 1990, 5(2):197–227.
[8] I. Guyon, A. Elisseeff, "An introduction to variable and feature selection". J. Mach. Learn. Res. **3**, 2003, pp1157–1182
[9] Olufunke.C. Olayemi, Olumide .S. Adewale, Olayemi. O. Olasehinde, Bolanle A. Ojokoh, Adebayo. O. Adetunmbi, (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). Federal University of Technology, Minna, Nigeria, held between 4th to 6th Sept ,2018
[10] M. Marlais, J. Evans .and E. Abrahamson “Clinical Predictors of admission in infants with acute bronchiolitis” May, 2011
[11] L. Todorovski and S. Dzeroski "Combining Multiple Models with Meta Decision Trees. Proceedings of the Fourth European Conference on Principles of Data Mining and Knowledge Discovery, Springer, Berlin, Germany, 1997 pp 54-64
[12] J. R. Quinlan, Book Review: C4.5: Programs for Machine Learning. San Francisco: Morgan Kaufmann. Publishers, Inc..2003
[13] Olasehinde O.,O., Alese B. K. and Adetunmbi A. O. " Stacked Ensemble of Intrusion Detection Systems with Multiple Model Tree Meta Algorithms" Paper accepted for presentation at the IEEE Cyber Security 2019, conference to be held at university of Oxford, Uk between 3rd and 4th of June, 2019.
[14] P. Yildirim, (2015). Filter-Based Feature Selection Methods for Prediction of Risks in Hepatitis Disease. International Journal of Machine Learning and Computing 5(4): 258 – 263.
[15[Coico,F. G ,Sunshine .G and Benjamin, Yildirim, P. (2015). Filter-Based Feature Selection Methods for Prediction of Risks in Hepatitis Disease. International Journal of Machine Learning and Computing 5(4): pg 258 – 263.
[16] M.A. Hall (1999). Correlation-based Feature Selection for Machine learning. PhD Thesis of the University of Waikato, Hamilton, New Zealand.

Author's Biography

Olayemi. O. Olasehinde obtained B.Tech, M.Tech and PhD. degrees in computer science in 1995, 2012 and 2018 respectively from the Department of Computer Science, Federal University of Technology, Akure, Nigeria. He also earned A Master degree in Business Administration (MBA) from the School of Management Studies of the same University in 2010. He has several publications in reputable peer reviewed journals and conference proceedings. His Research/Areas of Interest includes Information Security, Data Mining, Machine Learning, Bioinformatics and Behavioural Analysis, his current research is on Cyber Security and Privacy Protection .Dr, Olasehinde is a professional member of Nigeria Computer Society (NCS), Computer Professional Registration Council of Nigeria (CPN), Professional Statisticians Society of Nigerian (PSSN), full member

of the Institutes of Entrepreneurs of Nigeria (IOE). He is a Lecturer in the Department of Computer Science, Federal Polytechnic Ile-Oluji, Ondo State, Nigeria.

Olufunke Olayemi obtained B.Sc. in Computer sciences from University of Ado-Ekiti, Ekiti State Nigeria in 2002, . She earned her Master of Technology, M.Tech. and PhD. degrees in computer science in 20012 and 2018 respectively from the Department of Computer Science, Federal University of Technology, Akure, Nigeria. She has several publications in reputable peer reviewed journals and conference proceedings. Her Research/Areas of Interest includes Data Mining, Machine Learning and Bioinformatics and Health predictive models, her current research is development of predictive model for Upper Respiratory Tract Infections diagnosis. Dr. Olayemi is a professional member of Nigeria Computer Society (NCS), Computer Professional Registration Council of Nigeria (CPN), Nigeria Women in Information Technology (NIWIIT), She is a lecturer and researcher at the Department of computer Science, Joseph Ayo Babalola University, Ikeji-Arakeji, Osun State, Nigeria