Stacked Ensemble Model for Hepatitis in Healthcare System

Akinbohun Folake[1], Akinbohun Ambrose (Dr)[2], Oyinloye Oghenerukevwe E. (Dr)[3]

[1]*Department of Computer Science, Rufus Giwa Polytechnic, Owo, Ondo State, Nigeria*
[2]*Department of Ear, Nose and Throat/Head and Neck, University of Medical Sciences Teaching Hospital, Akure, Ondo State, Nigeria*
[3] *Department of Computer Science, Ekiti State University, Ado-Ekiti, Ekiti State, Nigeria*

***Abstract** -- Hepatitis is an inflammatory condition of the liver caused by a viral infection. Viral hepatitis is of various types namely A, B, C, D and E. Data and analytics driven models can be applied in medical domain with the aid of machine learning to predict diseases. The increase of the epidemiology of hepatitis needs computational intelligent tool for prediction. The objective of this paper is to develop a stacked ensemble model for hepatitis. The paper considers two feature selection methods namely correlation and consistency methods on the whole hepatitis dataset obtained from UCI repository. The Stacking ensemble method was selected which combined multiple classifications namely Decision tree (C4.5) and Naive Bayes (the base-level classifiers) via a meta-classifier namely classification via regression where cross validation was applied. On the level of ensemble learning, when classification via regression was used at meta-level on the reduced dataset, the result indicated that correlation method in a stacked ensemble model produces better prediction for hepatitis than consistency method. Correlation method on Decision Tree model can be used for prediction of hepatitis*

Keywords — *Consistency, decision tree, hepatitis, correlation*

I. INTRODUCTION

The inflammation of the liver cells (hepatocytes) is known as hepatitis. Hepatitis is an inflammatory condition of the liver that can be caused by a viral infection. Hepatitis could also be caused by autoimmunities, drugs, toxins, and alcohol. Viral hepatitis is of different types such as A, B, C, D and E [17]. The liver is the seat of metabolism of the body that performs many functions such as: bile production, filtering of toxins from your body, excretion of bilirubin (a product of broken-down red blood cells), cholesterol, hormones, and drugs and breakdown of carbohydrates, fats, and proteins etc. The moment the cells of the liver are inflamed, their functions become compromised. The severity of inflammation of the liver cells is largely dependent on the type of viral infection or the type of non viral causes of hepatitis. In viral infection of the liver (viral hepatitis), prognosis is dependent not only on the type of virus but also the presence of the following: ascites (extraperitoneal fluid meaning fluid outside the intestines in the abdominal cavity), increasing serum bilirubin, reducing serum albumin advanced age, marked fatigue, varices, associated spleenomegaly (meaning enlargement of the spleen), and presence of comorbidity (associated disease i.e. apart from the primary disease). Symptoms may not occur until the damage affects liver function that show some signs and symptoms of acute hepatitis which include the following: Dark urine, abdominal pain, flu-like symptoms, fatigue, loss of appetite, unexplained weight loss, yellow skin and eyes, which may be signs of jaundice [18]

Having considered the features and symptoms of hepatitis, clinical decision with the aid of machine learning methods can be complemented. Predictive data mining and ensemble methods are tools of machine learning. Data can be mined with predictive modeling to discover patterns in medical data to predict disease. With the large amount of data stored in databases and other repositories, it is important to develop a powerful tool for analysis and interpretation of such data and for the extraction of knowledge that could help in decision-making [2] [9]. Hence, data mining tools are assisting in producing result for disease diagnosis. The applications of data mining are widely used in classification and prediction technology in the field of bioinformatics. Predictive data mining tasks come up with a model from the available data set that is helpful in predicting unknown or future values. Ensemble systems have proven effective in computational intelligence and machine learning. Ensemble methods are developed to reduce bias, [16] variance and thereby improve the accuracy of an automated decision-making system.

II. REVIEW OF RELATED WORK

The application of learning algorithms on data brings out patterns and relationship which predicts future outcome for complementing clinical decisions. The burden of hepatitis has caused increased mortality and morbidity. According to World Health Organisation (2011) [14], about 130–170 million people are chronically infected with hepatitis C virus (HCV) and at risk of developing liver cirrhosis and/or liver cancer.

Jennifer *et al* (2015) [5] put up in their paper that in the Global Burden of Disease Study 2010, HBV was estimated to have resulted in 786,000 deaths, the vast majority being attributable to liver cancer (341,000 deaths) and cirrhosis (312,000 deaths).

The classification algorithms of various decision tree types were deployed in diagnosing hepatitis disease was proposed by Sathya (2011) [13]. C4.5 algorithm, ID3 algorithm and CART algorithm were used and the highest classification accuracy was of 83.2%.

Pinar (2015) [10] worked on a title: filter based feature selection methods for prediction of risks in hepatitis disease. The researcher used filter based feature selection viz: Cfs, principal components, Consistency, Info Gain, OneR Relief. Each was used on J48, IBK, Decision table and naïve bayes. The study showed that Naïve Bayes with Consistency, Information Gain, OneR and Relief methods performed better results than the others. Ensemble method was not deployed in the work.

A data mining approach for the prediction of hepatitis c virus protease cleavage sites was proposed by Ahmed (2011) [1] where the author tried to achieve more accurate prediction results, and more Informative knowledge about the HCV protein cleavage sites using Decision tree algorithm.

Yaming (2018) [15] applied a prediction model for Hepatitis B incidence in order to provide a scientific basis for Hepatitis B early detection and timely remediation. The researchers utilized principal component analysis to extract salient information, established a prediction model for Hepatitis B trends using the stepwise regression method, time series model and search data model for comparison and to predict the incidence of the next period. The results showed that the model provides stable and timely data.

Saranya and Seenuvasan (2017) [12] surveyed some data mining techniques to predict the liver disease. The study analyzed algorithms such as C4.5, Naive Bayes, Decision Tree, Support Vector Machine, Back Propagation Neural Network and Classification and Regression Tree Algorithms. Based on the conclusion of the study, it was seen that C4.5 gave better result compared to other algorithms.

Nuanwan *et al* (2007) [8] proposed a study on knowledge discovery for diagnosis using hepatitis C virus where data abstraction algorithm and the pruning algorithm were involved in order to extract a set of interesting rules. The outcome of the work produced useful rule set for physicians which assist to diagnose the hepatitis C virus (HCV) patient.

Roslina and Noraziah (2010) [11] worked a prediction of hepatitis prognosis. The methods used were wrapper method and support vector machine. The wrapper method was used to remove the bias and support vector machine was used for classification. The limitation of this work was that one feature method was used only one classification algorithm was deployed.

Nancy *et al* (2017) [6] investigated the application of feature selection on classification algorithms using hepatitis data. The methods applied three feature selection algorithms namely Fisher filtering, Relief filtering, Step Disc on 15 classification algorithms viz. Random Tree, Quinlan decision tree algorithm (C4.5), K-Nearest Neighbor algorithm etc., on a large hepatitis dataset (derived from the UCI Machine Learning Repository) that comprises of 20 attributes (including class) and 155 instances. The authors investigated on the importance of feature selections and classified the dataset using 15 most common classifiers. The results of this study indicate the level of accuracy as well as the importance of all the instances in detecting the survival of a person in future.

A paper developed by Mohammad *et al* (2016) [17] compared common methods such as decision trees, neural networks and Support Vector Machine (SVM) for Hepatitis diagnosis. The hepatitis dataset on uci repository was used. The result showed that neural network algorithm enjoyed highest accuracy of 89.74% in comparison with other algorithms.

III. METHODOLOGY

In order to achieve the stacked ensemble model for predicting hepatitis, there are stages namely pre-processing, feature selection stage, stacking ensemble method (base level construction and meta level construction) that are involved as presented in Figure 1.

A. Pre-processing stage

The hepatitis dataset was obtained from UCI repository. The dataset consists of 155 instances and 20 attributes (including class). The dataset contains missing values which were replaced. In order to handle the missing values, imputation methods such as median imputation, mean imputation and mode imputation can be used to fix the missing values.

B. Dataset of the hepatitis

Dataset comprises of the class and features. The features are: sex, age, varices, ascites, spleen palpable etc and the class label are live or die as presented in Table I

TABLE I
ATTRIBUTES INFORMATION

S/N	Feature/predictors	Predictors type
1	Age	Numeric
2	Sex	Male/female
3	Steroid	No/yes
4	Antiviral	No/yes
5	Fatigue	No/yes
6	Malaise	No/yes
7	Anorexia	No/yes

8	Liver big	No/yes
9	Liver firm	No/yes
10	Spleen palpable	No/yes
11	Spider	No/yes
12	Ascites	No/yes
13	Varices	No/yes
14	Bilirubin	Numeric
15	Alk phosphate	Numeric
16	Spot	Numeric
17	Albumin	Numeric
18	Protime	Numeric
19	Histology	No/yes
	Class	Live / Die

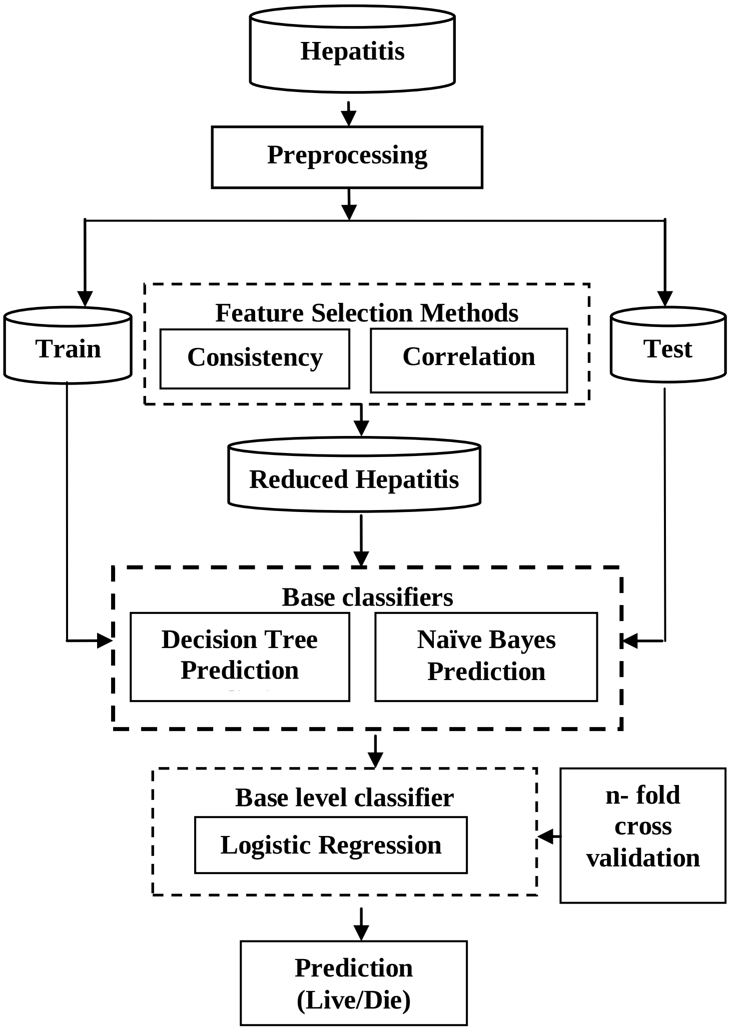

Figure 1: Architecture of the Proposed Stacked Ensemble for Hepatitis

C. Feature selection stage

After data pre-processing, there is need to use feature selection methods to remove redundant and irrelevant features. Feature selection performs important role in machine learning in order to build a robust model for either learning or classification from a large amount of data [3]. It is a phase that selects relevant features in the dataset without affecting the classification accuracy but improves accuracy. There are many algorithms that select important features/predictors but this work deploys two feature selection methods: correlation and consistency.

D. Stacking Ensemble method

Stacking was selected for this paper which combined multiple classifications (the base-level classifiers) via a meta-classifier. To use stacking ensemble method, various learning algorithms L_1, L_2, L_3,…L_n are combined on a single dataset, D which consists of D = (x_i,y_i).

In stacking ensemble learning, a phase uses a set of base-level classifiers C_1, C_2, …C_n which are generated where each classifier is learnt on the dataset i.e. $C_n = L_n$ (D).

- ***Base Level Classifier***

This paper uses two classification algorithms namely Decision tree (C45) and Naïve Bayes as the base classifiers.

1). Construction of Decision Tree (C4.5):

A decision tree is used as a classifier for determining an appropriate action for a given case [4]. For a patient who has hepatitis disease, it predicts whether the patient will die or live. 70% of the instances are used for training and 30% of the instances are used for test data. To find the appropriate class for a given patient (a person), it starts with the test at the root of the tree and keep following the branches as determined by the values of the features of the case at hand, until a leaf is reached.

The Decision Tree is constructed from the training sample (hepatitis) by splitting into subsets using greedy algorithm. Decision Tree (c4.5) uses information gain where entropy for each branch is calculated (i.e. the entropy of the class and each subset of the attribute/feature) are given in Equation 1. The algorithm (C4.5) uses Gain Ratio which is computed using Equation 2

$$E = -\sum_{i=1}^{n} p_i \log_2 p_i \qquad (1)$$

Where Pi is the proportion of examples in hepatitis that belong to the i-th class, n is number of classes.

$$\text{Gain Ratio}_{\text{attribute}} = \frac{\text{Gain}_{\text{attribute}}}{\text{SplitInfo}_{\text{attribute}}} \qquad (2)$$

2). Construction of Naive Bayes

Naïve Bayes classifier is a statistical classifier based on independency and probability (Bayes theorem). Naïve Bayes algorithm treat all features independently, with no feature depends on others features values.

Let k_{ij} be the hepatitis dataset containing records of i number of attributes, for j number of instance in the dataset such that, k_i is the set of attributes. $k = k_1,....,k_j$ are the predictors in the dataset. C is the class label for each predictors, C comprises of two classes is given as follows:

$$p(c_i \mid k_1, \ldots, k_j) = \frac{p(C_i)p(k|C_i)}{P(k_1, \ldots, k_n)} \qquad (3)$$

$where\ i = 0\ or\ 1$

Maximum posterior probability for classifying the class of a hepatitis instance is given as:

$$C= \underset{c}{argmax}\ p(C) \prod_{i=1}^{j} p(k_i \mid C_i) \qquad (4)$$

Naive Bayes will predict live or die with the highest probability.

- *Meta Level Classifier*

The predictions from the base classifiers are combined. These predictions were passed to the meta dataset and a meta algorithm was used. The next phase makes use of classification via regression algorithm at the meta-level classifier to be learnt which combines the outputs of the base-level classifiers. At the meta-level classification, the meta-level dataset consists of the form: $\bar{Y}^1_{n,}$ ...$\bar{Y}^m_{n,}$ a training set is generated where cross validation step is applied. In the procedure for cross validation, each of the base-level algorithms is applied to almost the entire dataset leaving one part for testing: C^i_n = L_n (D-d_i). The learned classifier is used to generate prediction.

IV. RESULTS AND DISCUSSION

When the methods have been setup, the result is presented below: Consistency method selected the following features: age, sex, malaise, spiders, ascites, varices, bilirubin, albumin, protein and histology. Correlation feature selection method selected 12 features namely: Ascites, Albumin, Bilirubin, Spiders, Varices, Malaise, Histology, Protein, Fatigue, Spleen Palpable, Age and Sex. The values that are less than 0.17 are cut-off.

The results of the two base classifiers/models namely Decision Tree (C45) and Naïve Bayes in the proportion of 70% training data and 30% of test data on the reduced dataset when using consistency method is presented in Table II. Results of using Correlation method on the Decision Tree (C45) and Naïve Bayes are shown in Table III

TABLE III
PERFORMANCE METRICS OF CONSISTENCY ON DECISION TREE (C45) AND NAIVE BAYES

Performance metric	**Decision tree (C45)**	**Naïve Bayes**
Accuracy (%)	80.43	80.43
Average Precision	0.782	0.852
Average Recall	0.804	0.804
F1 Score	0.789	0.819
Kappa statistic	0.2887	0.4864
Mean absolute error	0.2311	0.2114
Root mean squared error	0.3955	0.4076

TABLE IIIII
PERFORMANCE METRICS OF CORRELATION ON DECISION TREE (C45) AND NAIVE BAYES

Performance metric	**Decision tree (C45)**	**Naïve Bayes**
Accuracy (%)	84.78	82.61
Average Precision	0.842	0.883
Average Recall	0.848	0.826
F1 Score	0.844	0.840
Kappa statistic	0.4953	0.5588
Mean absolute error	0.2117	0.1823
Root mean squared error	0.3361	0.3786

Table IV showed the result of the meta level classifiers namely classification via regression in 10-fold cross validation performed on the reduced dataset using consistency method and correlation method.

TABLE IV
PERFORMANCE METRICS OF CORRELATION AND CONSISTENCY METHOD USING STACKED ENSEMBLE CLASSIFIER (CLASSIFICATION VIA REGRESSION)

Performance metric	**Correlation**	**Consistency**
Accuracy (%)	87.10	84.52
Average Precision	0.874	0.842
Average Recall	0.871	0.845
F1 Score	0.872	0.843
Kappa statistic	0.6151	0.5163
Mean absolute error	0.1963	0.2121
Root mean squared error	0.3216	0.3263

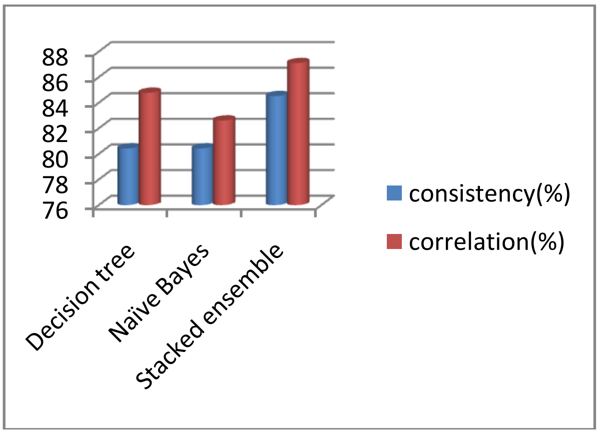

Figure 2: Accuracy of correlation and consistency methods on the classifiers

V. DISCUSSION ON THE PERFORMANCE METRICS OF HEPATITIS EVALUATION

The results of the base classifiers/models of Hepatitis using learning algorithms: Decision Tree (C4.5) and Naïve Bayes are discussed: Decision tree and Naïve Bayes had the same accuracy; and Naïve Bayes had higher F1 Score than Decision tree. The application of consistency method using Naïve Bayes model showed that the Naïve Bayes model is better than decision tree in predicting whether a patient who has hepatitis will live or die.

When the dataset was trained with correlation method using models namely Decision Tree and Naïve Bayes, the results showed that the Decision tree was higher in accuracy. Decision tree (C45) had higher F1 Score than Naïve Bayes. The application of correlation method showed that decision tree model had both better accuracy and F1 score.

On the level of ensemble learning, when classification via regression classifier was used at meta level on the reduced dataset, correlation produced 87.10% accuracy and F1 score of 0.872 while Consistency method gave accuracy of 84.52% with F1 score of 0.843. Looking at all parameters, the result indicated that using correlation method in a stacked ensemble model produces good prediction.

VI. CONCLUSION

One effective way of analyzing this data is through prediction by using machine learning tools. Using feature selection methods on models assist in remove noise from the data. Choosing of better base models is due to the size and structure of the dataset obtained. Stacked ensemble models can be used in predicting the class of hepatitis in patients which improves prediction accuracy. In health care system, algorithmic tools predict which people will live or die when features of hepatitis set-in. Hence ensemble systems have come a long way in complementing the clinical decision.

REFERENCES

[1] Ahmed Mohamed Samir Ali Gamal Eldin (2011). A data mining approach for the prediction of hepatitis c virus protease cleavage sites. International Journal of Advanced Computer Science and Applications (IJACSA),vol. 2, No. 12. Page 179-182

[2] Fayyad U. M., Piatetsky-Shapiro G., Smyth P., and Uthurusamy R. (1996). Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press

[3] Guyon I. and Elisseeff A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research, vol. 3, pp. 1157-1182

[4] Han J, Kamber M, Pei J (2012). Data mining: concepts and techniques, 3rd edition. Elsevier, Amsterdam

[5] Jennifer H. MacLachin and Benjamin C. Cowie (2015). Hepatitis B Virus Epidemiology. Cold spring harb perspect med

[6] Nancy P., Sudha V. and Akiladevi R. (2017). Analysis of feature Selection and Classification algorithms on Hepatitis Data. International journal of advanced research in computer engineering & technology (IJARCET). Volume 6, Issue 1

[7] Mohammad Reza Moshkani, Mahdi Rousta and Yaghoub Farjami(2016). Eamining and comparing data mining-based techniques for hepatitis diagnosis. International journal of innovative research in computer and communication engineering, vol. 4, issue 1, page 750-756

[8] Nuanwan Soonthornphisaj, Supakpong Jinarat, Taweesak Tanwandee and Masayuki Numao (2007). Knowledge Discovery for Hepatitis C Virus Diagnosis: A Framework for Mining Interesting Classification Rules. International conference on intelligent computing. Advanced intelligent computing theories and applications with aspects of contemporary intelligent computing techniques. Volume 2, Page 171-179

[9] Piatetsky-Shapiro G. and W. J. Frawley W. J. (1991). Knowledge Discovery in Databases. AAAI/MIT Press.

[10] Pinar Yildirim (2015). Filter Based Feature Selection Methods for Prediction of Risks in Hepatitis Disease. International Journal of Machine Learning and Computing, Vol. 5, No. 4 Page 258-263

[11] Roslina A.H., and Noraziah A. (2010). Prediction of hepatitis prognosis using support vector machine and wrapper method. IEEE, 2209-2211

[12] Saranya A. and Seenuvasan G. (2017). A comparative study of diagnosing liver disorder disease using classification algorithm. International Journal of Computer Science and Mobile Computing. (IJCSMC)Vol. 6, Issue. 8, pg.49 – 54

[13] Sathya Devi G., (2011). Application of CART algorithm in hepatitis disease diagnosis. IEEE, 1283-1287

[14] World health organization media centre. “Hepatitis C." http://www.who.int. 2011. 5 October 2011

[15] Yaming Zhang, Yang Zhao, Xin Lin, Aibo Wang Dandan Che (2018). Modeling for the prediction of Hepatitis B incidence based on integrated online search indexes. Informatics in Medicine unlocked.volume 10, pages 143-148

[16] Zhi-Hua Zhou (2012). Ensemble Methods: Foundations and Algorithms. CRC Press

[17] Available : http://www.healthline.com/

[18] Available: http://www.medicalnewstoday