Abstract—Head and neck cancers (HNC) are indicated when cells grow abnormally. The incidence of HNC is on the increase owing to several factors. There is often late presentation that can result in loss of lives (mortality) especially in Africa due to paucity of specialists. These challenges prompted the development of a stacked ensemble model for diagnosis of HNC to facilitate prompt referral. The data were collected which consists of 1473 instances with 18 features. Information Gain was used for selecting important features and three supervised learning algorithms were deployed for the base learners: Decision Tree (C4.5), K-Nearest Neighbors and Naïve Bayes. The predictions of the base learners were combined and passed to meta learners: Logistic Model Tree (LMT). The result showed that Information Gain method with stacked LMT produced higher accuracy than that of the base learners’ results. Hence, this stacked model can be used for diagnosis of HNC in healthcare systems.

Index Terms—Decision Tree, Larynx, Naïve Bayes, Sinonasal.

I. INTRODUCTION

Head and neck cancer (HNC) are the types of cancer which occur as a result of uncontrollable growth of abnormal cells in head and neck regions of the body. Cancers can affect any part of human body but if they affect head and neck regions, they are known as head and neck cancers (HNC). The part of human body that HNC affect could be of various parts in the head and neck regions. These are: nasopharyngeal, thyroid, larynx, sinonasal etc. There are factors such as chemical or toxic compound exposure, ionizing radiation, pathogens, human genetics and so on that can predispose to HNC.

Patients with HNC present some symptoms which include weight loss, bleeding, swelling, hemoptysis, dyspnoea etc. Fatigue is a significant symptom of head and neck cancer patients [1]. These (symptoms) and signs are what we called features of HNC. The features were considered to predict the types of cancer around head and neck regions.

II. RELATED WORK

Some works by researchers which included the use of machine learning algorithms to diagnose diseases have been

DOI: http://dx.doi.org/10.24018/ejers.2020.5.9.2095
reviewed. They are, among others:

Two prediction models (Radiotherapy (RT) and End of Treatment (EOT)) at different time points were developed to predict weight loss ≥5 kg (yes/no) at 3 months post-Radiotherapy (RT) using Classification and Regression Tree (CART) algorithm [5]. The study indicated that CART could be employed for weight loss prediction in patients with head and neck cancer.

[6] evaluated the use of Classification and Regression Tree (CART) model in predicting weight loss following head and neck cancer radiation therapy. Two prediction modes were developed to predict weight loss ≥5 kg at 3 months’ post-radiation therapy. The modes were (1) during radiation therapy planning using patient demographic, delineated dose data, planning target volume–organs at risk shape relationships data and (2) at the end of treatment (EOT) using additional on-treatment toxicities and quality of life data. This showed that CART was deployed to predict weight loss in head and neck cancers patients.

[7] evaluated a pattern of follow up visits among patients with head and neck cancers in Jos, North-Central, Nigeria. Data were collected from Jos University Teaching Hospital, Jos, North-central Nigeria and analyzed using the statistical software.

[8] considered a classifier model using Machine Learning Algorithms on differential diagnosis of suspicious thyroid nodules via Sonography (Ultrasound). The study included 970 histopathologically proven thyroid nodules in 970 patients. Two radiologists retrospectively reviewed ultrasound images, and nodules were graded according to a five-tier sonographic scoring system. Naïve Bayes, Radial Basis Function Neural Network, Support Vector Machine were employed for diagnosis of thyroid nodules.

[9] described competing causes of death in the head and neck cancer population. They identified patients with first mucosal head and neck cancer (HNC) from the Surveillance, Epidemiology and End Result database. The study showed that patients with head and neck cancer died from competing causes.

[10] developed a feature selection method for oral cancer using Apriori algorithm where the original algorithm of Boolean association rules of mining frequent item sets was employed. The data mining methods were explored to identify the suitable techniques for efficient classification of data.

[11] deployed Back Propagation Algorithm Neural Network to diagnose thyroid disease. The study indicated that developed neural network could be used as a diagnostic tool for earlier prediction of a thyroid disease.

[12] worked on the prediction of acute myeloid leukemia cancer using data mining. Data mining techniques such as Bayes Network, Jrip, J48, Multilayer Perceptron, IBK, Decision Tree were used on the dataset. The study showed that data mining algorithms could be used for myeloid leukemia cancer prediction.

[13] carried out a study where the objective of the study was to predict breast cancer survivability where different data mining techniques: Naïve Bayes, Back-Propagated Neural Network, and the C4.5 decision tree algorithms were adopted. It was established that C4.5 algorithm had a better performance than the other two techniques. The study was limited to only one type of cancer (breast).

Prediction of survival in patients with esophageal carcinoma (cancer) using Artificial Neural Networks was conducted by [14]. It was observed that accurate estimation of outcome in patients with malignant disease was an important component of the clinical decision-making process.

III. MATERIALS AND METHODS

The architecture of a stacked ensemble model to diagnose head and neck cancers is presented in Figure 1.

![Figure 1. Architecture of a proposed model for HNC diagnosis](image)

NB represents Naïve Bayes; DT means Decision Tree; KNN is the K-Nearest Neighbors and LMT is the Logistic Model Tree; HNC connotes Head and Neck Cancer.

The major components of the architecture for HNC diagnosis model are collection of data, preprocessing, feature selection, base models and meta models/super learners.

The components of the proposed ensemble model for diagnosis of head and neck cancers are explained as follows:

A. Collection of Data

Data were collected from three Hospitals in Nigeria: Obafemi Awolowo University Teaching Hospital, Ile-Ife; Federal Medical Centre, Owo and University of Medical Sciences Teaching Hospital, Akure. The data were preprocessed. Information Gain was deployed for feature selection so that irrelevant feature could be removed from the dataset. The HNC dataset gathered consists of 18 features which are: bleeding, poor appetite, weight loss, snoring, swelling, nasal blockage, mouth breathing, hypoponasal speech (poor speech), halitosis, facial asymmetry, fatigue, hoarseness, dyspnea, tinnitus, haemoptysis, proptosis, odynophagia and dysphagia. The
classes include: Sinonasal, nasopharyngeal, laryngeal and thyroid. The number of instances is 1473. The attribute values are converted to numeric.

B. Feature Selection

During data collection, problems like irrelevant features might occur. We introduced feature selection method where Information Gain algorithm was used to remove irrelevant features from the dataset (HNC), to improve accuracy and to reduce training time. This task was tackled in the prepared data before applying learning algorithms. The Information Gain as a feature selection method is described thus:

Information Gain is used to measure the dependence between features and labels and calculates the information gain between the feature and the class labels C. It is a method that ranks features based on a relevancy score which is based on each individual attribute [15].

It finds the entropy of the classes, features subsets and weighted average entropy in order to get the information gain of each feature. The weighted information gains of the features are ranked. In information gain, a feature is relevant if it has a high information gain.

To calculate the information gain, the entropy of the HNC type (class) is calculated using Equation 1.

\[ E = -\sum_{i=1}^{n} P_i \log_2 P_i \]  

(1)

Where \( P_i \) is the proportion of examples in HNC that belongs to the i-th class, \( n \) is number of classes and \( E \) is the entropy. Information Gain is expressed as the difference between the prior entropy of classes and posterior entropy [16].

C. Construction of the base model

The base models deployed for this study were Decision Tree, K-Nearest Neighbor and Naïve Bayes. The dataset was divided into: training data and testing data. 1031 records were used for training data while 442 records were used for test data. The training data or cases were assumed to be represented as a pair \([x_1, x_2, x_3, \ldots x_n \rightarrow y]\) where \( x_1, x_2, x_3 \ldots x_n \) are vectors of attribute values describing some cases while \( y \) is the appropriate class or target. The fist base model is explained as follows:

A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences. Decision tree is a predictive model to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). Decision tree is constructed in a top-down recursive divide-and-conquer manner [17].

The goal of decision tree is to create a model that predicts the value of a target variable based on several input variables. A tree can be "learned" by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called recursive partitioning. It is a flow-chart-like structure, where each internal (non-leaf) node denotes a test on an attribute, each branch represents the outcome of a test, and each leaf (or terminal) node holds a class label. These nodes in turn will continue to split until a final node or leaf node is grown. The leaf node determines the final classification of the variable being tested.

The second base model employed in this paper was K-Nearest Neighbors (KNN).

The k-NN algorithm belongs to the family of instance-based and lazy learning algorithms [18]. Instance-based algorithms are those algorithms that model the problem using data instances (or rows) in order to make predictive decisions [19].

When new unlabeled data comes in, k-NN operates in two basic steps: Firstly, it looks at the k closest labeled training data points. Secondly, using the neighbors’ classes, k-NN gets a better idea of how the new data should be classified. For continuous data, k-NN uses a distance metric like Euclidean distance to locate what is closer to k [20]. The Euclidean distance is given in Equation 2

\[ \text{Euclidean} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \]

(2)

where \( x \) represents the query-instance and \( y \) represents all the training samples;

\( (x_i - y_i) \) represents distance between the query-instance and all the training samples.

The third base model was Naïve Bayes classifier. It is a classification technique and frequently trumps a more sophisticated predictive analytic tool which is designed for use when predictors are independent of one another within each class [21].

Naïve Bayes Classifier technique is based on Bayesian theorem. It is a conditional probability model that enables conditional predictions. Bayes theorem provides a way of calculating posterior probability \( P(c|x) \) from \( P(c) \), \( P(x) \) and \( P(x|c) \) using the Equation 3:

\[ P(c|x) = \frac{P(x|c)P(c)}{P(x)} \]

(3)

\( P(c|x) \) is the posterior probability of class (target) given predictor (attribute) of class.

\( P(c) \) is called the prior probability of class.

\( P(x|c) \) is the likelihood which is the probability of predictor of given class.

\( P(x) \) is the prior probability of predictor of class.

It combines this model (conditional probability) with a decision rule. One common rule is to pick the hypothesis that is most probable; this is known as the Maximum a Posteriori (MAP) decision rule.

D. Stacking with Logistic Model Tree (LMT)

The predictions of the base models were combined, and at super level or meta level, Logistic Model Tree Algorithm was used for building the stacking.

Logistic Model Tree (LMT) is a supervised training model that combines Logistic Regression and Decision Tree [22].

Logistic model tree is based on a decision tree that has linear regression model at its leaves to provide a piecewise linear regression model. The algorithm makes use of cross validation to find a number of LogitBoost iterations that prevent overfitting.
IV. RESULT

When Information gain algorithm was applied, the threshold was set at <0.2, the result is presented in Table 1 where features with higher values were used and passed for the training of base models and a meta model.

### TABLE I: RESULT OF INFORMATION GAIN

<table>
<thead>
<tr>
<th>Ranked Feature</th>
<th>Information Gain Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nasal blockage</td>
<td>0.791</td>
</tr>
<tr>
<td>Dyspnoea</td>
<td>0.757</td>
</tr>
<tr>
<td>Haemoptysis</td>
<td>0.757</td>
</tr>
<tr>
<td>Snoring</td>
<td>0.749</td>
</tr>
<tr>
<td>Hoarseness</td>
<td>0.74</td>
</tr>
<tr>
<td>Facial Asymmetry</td>
<td>0.542</td>
</tr>
<tr>
<td>Fatigue</td>
<td>0.362</td>
</tr>
<tr>
<td>Hypopnasal</td>
<td>0.333</td>
</tr>
<tr>
<td>Speech</td>
<td>0.325</td>
</tr>
<tr>
<td>Proptosis</td>
<td>0.317</td>
</tr>
<tr>
<td>Tinnitus</td>
<td>0.30</td>
</tr>
<tr>
<td>Poor Appetite</td>
<td>0.30</td>
</tr>
<tr>
<td>Odynophagia</td>
<td>0.303</td>
</tr>
<tr>
<td>Mouth Breathing</td>
<td>0.302</td>
</tr>
<tr>
<td>Halitosis</td>
<td>0.269</td>
</tr>
<tr>
<td>Swelling</td>
<td>0.236</td>
</tr>
<tr>
<td>Bleeding</td>
<td>0.225</td>
</tr>
<tr>
<td>Dysphagia</td>
<td>0.206</td>
</tr>
</tbody>
</table>

The predictions of the three base models (Decision tree (C4.5), KNN and Naïve Bayes) were stacked with Logistic Model Tree. This was used on the selected features of Information Gain. The stacking was done on the basis of cross validation. The results are presented in Table 2 and accuracies of the base models on Information Gain is shown in Figure 2

### TABLE II: RESULTS OF THE PERFORMANCE EVALUATION FOR THE BASE MODELS WITH INFORMATION GAIN

<table>
<thead>
<tr>
<th>Base models</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision Tree</td>
<td>93.21%</td>
<td>0.943</td>
<td>0.932</td>
<td>0.931</td>
</tr>
<tr>
<td>KNN</td>
<td>94.80%</td>
<td>0.949</td>
<td>0.948</td>
<td>0.948</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>94.12%</td>
<td>0.945</td>
<td>0.941</td>
<td>0.941</td>
</tr>
</tbody>
</table>

A. Results of the Information Gain with Decision Tree Data

The selected features by the information gain were trained on the base models: Decision Tree (C45), KNN and Naïve Bayes in the ratio 70% of the dataset for training set and 30% of the data for test set. The performance metrics we considered for base level are accuracy, precision recall and F1 score. The results are presented in Table 2 and accuracies of the base models on Information Gain is shown in Figure 2.

B. Stacking with Logistic Model Tree (LMT) with Information Gain Selection

The predictions of the three base models (Decision tree (C4.5), KNN and Naïve Bayes) were stacked with Logistic Model Tree. This was used on the selected features of Information Gain. The stacking was done on the basis of cross validation. The results is presented in Table 3

### TABLE III: RESULT OF THE PERFORMANCE EVALUATION OF STACKED ENSEMBLE MODEL USING LMT

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>95.11%</td>
</tr>
<tr>
<td>Precision</td>
<td>0.952</td>
</tr>
<tr>
<td>Recall</td>
<td>0.951</td>
</tr>
<tr>
<td>F1 Score</td>
<td>0.951</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.9336</td>
</tr>
<tr>
<td>Statistic</td>
<td></td>
</tr>
<tr>
<td>Mean Absolute</td>
<td>0.0304</td>
</tr>
</tbody>
</table>

V. DISCUSSION

From the results presented, the following facts are inferred:

The HNC dataset was trained on Information Gain method with base models namely Decision Tree, KNN and Naïve Bayes. The result showed that the accuracy of Decision Tree (C4.5) was 93.21%, KNN was with accuracy of 94.80% and Naïve Bayes with accuracy of 94.12%. Based on accuracy, it indicated that KNN had the highest accuracy of 94.80% followed by Naïve Bayes with accuracy of 94.12% while Decision Tree had the least accuracy of 93.21%. Using F1 score for evaluation, Decision Tree, KNN and Naïve Bayes have 0.931, 0.948 and 0.941 respectively. From this metric, it showed that KNN had the highest F1 Score followed by Naïve Bayes and the least was Decision Tree. This implied that KNN was a better predictive model to predict or diagnose the type of HNC when Information Gain method was applied in the ratio 70% of the dataset for training set and 30% of the dataset for test set.

Following the predictions of the base learners, they (predictions) were combined. Logistic Model Tree was used at a stacked meta level to train the HNC dataset in a stratified cross validation. This was done on the selected features using Information Gain.

The results of the stacked model of Head and Neck Cancer dataset was 95.11% for Logistic Model Tree (LMT). This showed that LMT was higher than the values of the three base models. i.e. There was an improvement of the accuracy.

VI. CONCLUSION

HNC has become a global disease that its mortality and morbidity rate are on the increase because of late presentation and lack of access to the specialists in Ear, Nose, Throat/Head and Neck. Due to the common clinical features patients with HNC present with, it is of importance to deploy the use of stacked ensemble learning algorithms to predict the appropriate head and neck cancer type.

This study recommends that machine learning algorithms can be used to diagnose head and neck cancers early in developing countries especially as patients in these countries (developing) present to primary health care facilities as first port of call. The study helps primary health workers to diagnose early and refer to the tertiary health institutions where ENT specialists are found.

---

**Fig 2. Accuracy of the base models with Information Gain**
ACKNOWLEDGMENT

The following people are acknowledged: Prof Esan (OAUTH, Ife), Dr Pelemo (UNIMEDTH, Akure), Dr Ariyibi (FMC, Owo), Dr Ameye (OAUTH, Ife)

REFERENCES


Folake Akinbobun works with Rufus Giwa Polytechnic, Owo, Ondo State, Nigeria. She has Master of Technology in Computer Science from Federal University of Nigeria in 2012. She has rounded up her Ph.D in Computer Science at the same University. Her areas of specialization are Data Science and Machine Learning. She has many publications both at local and international levels. She is a member of Nigeria Computer Society (NCS) and Women in Artificial Intelligence (WAI).

Ambrose Akinbobun is Principal Consultant in Ear, Nose, Throat, Head and Neck Surgery of University of Medical Science Teaching Hospital, Akure, Nigeria. He possesses special interest in Head and Neck Surgery. He is a trainer of medical students and Residents (Post Graduate Medical Doctors in Training). He is the current Head of Department and has authored a number of publications. He is a public speaker at both radio and television programmes. He is a fellow of the West African College of Surgeons and an examiner of the West African College of Surgeons.

Adekanle Daniel is a Senior Lecturer and Honorary Consultant, Department of Otorhinolaryngology, Head and Neck, University of Ibadan, Oyo State, Nigeria. He has a number of peer-reviewed publications to his credit.

Oyiynlo Ogheneurewu Elohor is a lecturer in Computer Science Department, Ekiti State University, Ado Ekiti, Ekiti State, Nigeria. She bags PhD in Computer Science from Federal University of Technology, Akure, Nigeria. She is the Head of Department. She has many publications to her credit.