Application of Machine Learning for Diagnosis of Head and Neck Cancer in Primary Healthcare Organisation

Olatunbosun Olabode, Adebayo O. Adetunmbi, Folake Akinbohun and Ambrose Akinbohun

***Abstract*—Head and neck cancers (HNC) are indicated when cells grow abnormally. The disturbing rate of morbidity and mortality of patients with HNC due to late presentation is on the increase especially in Africa (developing countries). There is need to diagnose head and neck cancer early if patients present so that prompt referral could be facilitated. The collected data consists of 1473 instances with 18 features. The dataset was divided into training and test data. Two supervised learning algorithms were deployed for the study namely: Decision Tree (C4.5) and k-Nearest Neighbors (KNN). It showed that Decision Tree outperformed with accuracy of 91.40% while KNN had accuracy of 88.24%. Hence, machine learning algorithm like Decision Tree can be used for diagnosis of HNC in healthcare organizations**.

***Index Terms*—Decision Tree, Head and Neck Cancer, k-NN, Nasopharyngeal.**

I. Introduction

Head and neck cancers (HNC) are the sixth most common cancers [12] of more than one hundred cancer types world-wide. Cancer is the abnormal cell growth that can affect any parts of the body. When cancer (malignancy) affects the head and neck regions, excluding the eye and the brain, it is termed head and neck cancer.

There is a disturbing increase in incidence globally; especially in developing countries of the world. Head and neck regions of the body are conspicuously located anatomically and easily visible making abnormal growth in such part of the body a great cosmetic embarrassment.

Several predisposing factors to head and neck cancer include: exposure to carcinogens like chemical dust, cigarette smoke, viruses, iron deficiency, familiar risks and molecular factors.

Morbidity and mortality of patients with HNC in developing countries like Nigeria is on the increase due to late presentation at the tertiary health institution where medical specialists in such field can be found [5],[4]. Patients often present at the primary health centre where they are misdiagnosed and suffer from delay referral.

One of the tasks in machine learning is classification. The goal of classification is to accurately predict the target class for each case in the data. In the model (training) process, different classification algorithms use different techniques for finding relationships between the values of the features and the values of the class. Application of machine learning algorithm can be used to diagnose head and neck cancer type from the available data.

The need to deploy computer-based diagnostic tool to enhance prompt diagnosis at the primary health centre level cannot be overemphasized.

II. Related Work

Researchers who had worked on the burden, self-efficacy, predictors of head and neck cancers and other types of cancer using machine learning techniques are given below:

[1] underscored that the burden of head and neck cancers in Nigeria was evident. The occurrence, distribution, identified risks factors, presentations, diagnostic method, treatment, prognosis and challenges associated with the management of HNC are examined. The result showed that late presentation of the advanced head and neck was a common report from different parts of Nigeria.

Patients with Head and Neck Cancer often have facial disfigurement. Linear regression analysis was used to determine the associations between facial disfigurement and outcome variables like psychological distress, distress in reaction to unpleasant behavior of others, and social isolation. The study showed that the degree of facial disfigurement was positively related to psychological distress [7].

[2] proposed a model on prediction of survival of head and neck cancer's patients. Prognostic value of site of the primary tumor, age at diagnosis, gender, T-N- M-stage, and prior malignancies of 1396 patients were studied univariately by Kaplan-Meier curves and the log-rank test. The Cox-regression model was used to investigate the effect of these variables simultaneously on a prediction model of survival in individual patients.

[9] studied the clinical implications of malnutrition in patients with head and neck cancer during treatment. It was observed that regression analysis showed a significant association between weight loss and deterioration of global quality of life. The study established that there was association between weight loss and quality of life in head and neck cancer patients.

[15] proposed significant predictors like fatigue, global health/QOL, social contact, speech, pain, swallowing, and xerostomia were also identified as predictive of depression and quality of life (QOL) among long-term survivors of head and neck cancer. T-test and linear regression analyses

Published on April 26, 2020.

F. Akinbohun is with Rufus Giwa Polytechnic, Owo, Ondo State, Nigeria (e-mail: folakeakinbohun@yahoo.com). She is the corresponding author.

O. Olabode works with the Computer Science Department, Federal University of Technology, Akure, Ondo State, Nigeria (e-mail:oolabode@futa.edu.ng).

A. O. Adetunmbi lectures at the Department of Computer Science, Federal University of Technology, Akure, Ondo State, Nigeria (e-mail:aoadetunmbi@futa.edu.ng).

A. Akinbohun is with University of Medical Sciences Teaching Hospital, Akure, Ondo State, Nigeria (e-mail: akinbohunambrose@yahoo.com).

were used to construct predictive models.

[14] predicted acute myeloid leukemia cancer where data mining techniques such as Bayes Network, JRip, J48, Multilayer perceptron, IBK, Decision Tree were used on the dataset. Even though, they used many classification algorithms on the dataset, no feature selection method was performed for better performance.

Prediction of leukemia (blood cancer) was proposed by Rajeswari and Aruchamy [3]. The goal of the study was to survey data mining techniques to predict leukemia. The study surveyed techniques of classification algorithms such as Naïve Bayes, Artificial Neural Network and Support Vector Machine. They combined multiple models for gene expression analysis and filter methods such as gain ratio, relief-f, and support vector machine filter were used.

Prediction System of Larynx Cancer was presented by Benjamín and Carlos (2013). One component of the proposed prediction system was the transformation and selection of data; the second component was a set of classifiers to obtain the prediction of life of sample patients with this type of cancer.

III. Materials and Method

The use of machine learning algorithms has been a great tool for diagnosis of diseases. The application of machine learning in diagnosis of head and neck cancer contains various components which include collection of data, data preprocessing, applications of models on the dataset and result.

A. Collection of Data

The raw data were collected from three hospitals: University of Medical Sciences, Akure, Federal Medical Centre, Owo, Ondo State and Obafemi Awolowo University Teaching Hospital Complex, Ife, Nigeria. Data from Pathology and Ear, Nose, Throat/Head and Neck (ENT/H &N) Departments were harmonized.

B. Preprocessing

The collected records were made up of 1473 instances, 18 features/attributes and 4 classes which formed a dataset. The classes were nasopharyngeal cancer, sinonasal cancer, laryngeal cancer and thyroid cancer. Attribute values were converted to numeric for easy processing.

The dataset is made up of features and the attribute values in Table I. The classes of the dataset are described briefly below:

Thyroid cancer: Cancer that affects the front part (anterior) of the neck where the gland is located.

Nasopharyngeal cancer: Cancer that affects the air passage behind the nose.

Laryngeal cancer: Cancer that affects the voice box.

Sinonasal cancer: It occurs in the nasal cavity and paranasal sinuses.

TABLE I: HNC Dataset

Feature	Attribute
Bleeding	Mild/moderate/severe
Poor appetite	Yes/no
(Anorexia)	Yes/no
Weight loss	Yes/no
Snoring	Small/medium/large
Swelling	Yes/no
Nasal blockage	Yes/no
Mouth breathing	Yes/no
Hyponasal speech	Yes/no
Halithosis	Yes/no
Facial Asymmetry	Yes/no
Fatigue	Yes/no
Hoarseness	Yes/no
Dyspnoea	Yes/no
Tinnitus	Yes/no
Hemoptysis	Yes/no
Proptosis	Yes/no
Odynophagia	Sinonasal
Cancer type	/nasopharyngeal/ laryngeal/ thyroid

C. Construction of the Model

The dataset which was made up of 1473 instances was divided into two parts: training and test data. The training data constituted 70% of the dataset and 30% of the dataset were used for test data. The two supervised learning algorithms namely Decision tree (C45) and k-Nearest Neighbors (k-NN)) were applied for the purpose of classification task. The models/algorithms were employed for diagnosis of head and neck cancer as explained below:

1) Decision Tree

Decision tree is a predictive modeling approach that is used for both classification and regression problems. Decision tree is an algorithm based on tree structure where each node indicates a feature, each branch represents decision rule and each leaf depicts an outcome/target/class [10], [11].

A decision tree uses divide and conquer algorithm to split a node into two or more sub-nodes which is done in respect to outcome/class. On each iteration of the algorithm, it iterates the unused feature/attribute of the dataset and calculates Entropy(H) and Information gain (IG) of this feature/attribute. The feature with highest information gain forms the root node which is partitioned into other sub-nodes and tested with another feature recursively until the leaf is reached (outcome).

The equations of Entropy and Information Gain are given in Equations 1 and 2.

In order to calculate the information gain, the entropy of the feature/attributes and class are to be calculated using equations as given:

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

Where P_i is the proportion of examples in HNC that belongs to the i-th class
n is number of classes, E is entropy

$$\text{Gain} = \text{Info (HNC)} - \text{Info (Attribute)} \quad (2)$$

Algorithm 1: Decision tree

1. If S is labeled with the same class, it returns a leaf labeled with class (HNC type)

2. Choose test, t (criterion) that has two or more mutually exclusive outcomes as: {∂_1, ∂_2, ∂_3…∂_n} (i.e. which feature to pick for splitting).
3. Partition S into disjointed subset S_1, S_2, S_3…S_n such that S_i={i} for the test, t, for i = 1, 2, …n. To get best test that partitions the passed training HNC into subsets {S_1, S_2, S_3…S_n} such that each subset S_n contains cases that are of the same class.
4. There will be a call on each of the subsets S_1, S_2, S_3, … S_n) recursively
5. Let decision tree return by these recursive calls be C_1, C_2, C_3, … C_n)
6. The call on the entire training HNC returns the tree. (Adapted: [8])

Decision Tree to Decision Rules of the HNC Dataset

Decision tree is transformed to a set of rules by mapping from the root node to the leaf nodes one by one. Four features of HNC were used to present the decision rule of the Decision tree. Fig. 1 shows the decision tree of the HNC as follows:

R_1: IF (bleeding = mild) AND (fatigue = No) THEN Cancer type = Laryngeal cancer
R_2: IF (bleeding = mild) AND (fatigue = Yes) AND (hyponasal speech = Yes) THEN Cancer type = Nasopharyngeal cancer
R_3: IF (bleeding = mild) AND (fatigue = Yes) AND (hyponasal speech = No) THEN Cancer type = Sinonasal cancer
R_4: IF (bleeding= moderate) THEN Cancer type = Thyroid cancer
R_5: IF (bleeding = severe) AND (snoring = No) THEN Cancer type = Sinonasal cancer
R_6: IF (bleeding = severe) AND (snoring = Yes) THEN Cancer type = Nasopharyngeal cancer.

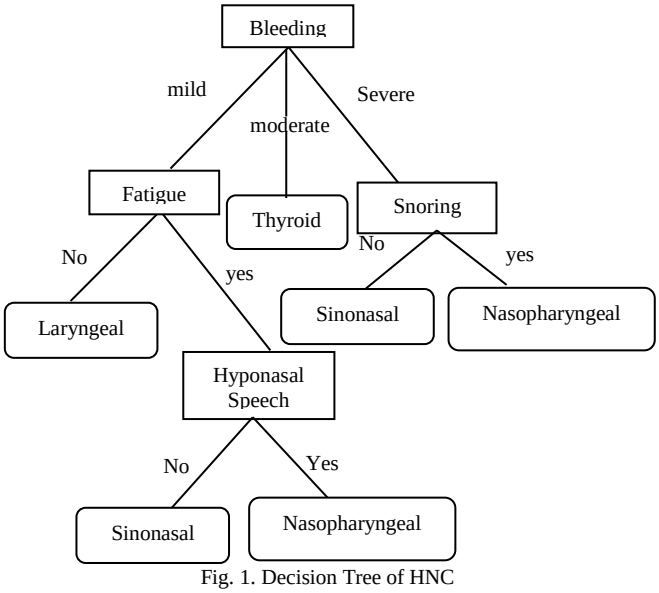

Fig. 1. Decision Tree of HNC

2) K-Nearest Neighbor

k-Nearest Neighbor (k-NN) is an algorithm that is used for classification and regression task [1]. It is used for predictive decision making. The algorithm works on the known dataset of which it contains unknown class or target that needs to be classified. k-NN assigns weights to the neighbors, so that the nearer neighbors give more to the average than the more distant ones. The neighbors are taken from a set of objects for which the class is known. This can be thought of as the training set for the algorithm. The unknown is classified by a simple plurality vote of its neighbor, where the class of close neighbours supersede [16].

For the purpose of this study, k-NN was used for classification task in which distance metric was used for continuous variables. The distance metric is computed by using Euclidean distance as stated in Equation 3. Algorithm 2 shows the step-by-step by which k-NN works

$$Euclidean = \sqrt{\sum_i^n (ID_{xi} - ID_x)^2} \quad (3)$$

Algorithm 2: K-Nearest Neighbor

Classify (X, C, x) // xi: Training data (features), C: Class labels (HNC types), // x: unknown sample

For i = 1 to n do
Compute Euclidean distance (x_i, x)
$Euclidean = \sqrt{\sum_i^n (ID_{xi} - ID_x)^2}$
Sort Euclidean distance
end for
set I containing indices for the k smallest distances d (x_i, x)
return majority label for {C, where i∈I)

IV. Results

After the methods had been set up experimentally, the results are produced which are given below: The features were passed in the HNC engine for models' construction using Decision Tree and K-Nearest Neighbours. The performance evaluation metrics for the models we considered were accuracy, precision, recall and F1 score. The results of Decision Tree and k-NN models were presented in Table II.

TABLE II: Results of the Performance Evaluation Metrics of the Models

Models	Accuracy	Precision	Recall	F1 Score
Decision Tree	91.40%	0.9258	0.8859	0.9054
KNN	88.24%	0.9002	0.8789	0.8894

V. Discussion

The results of the models (Decision Tree (C4.5) and KNN) are given below: The result showed that the accuracy of Decision Tree (C4.5) was 91.40% and KNN had accuracy of 88.24% respectively. It meant that Decision Tree had higher accuracy than KNN. F1 score is a metric that evaluates the harmonic mean of precision and recall and measures a test's accuracy; hence it is a good measure which determines the best model. From the results, the F1 score for Decision Tree was 0.9054, KNN had F1 score of 0.8894. This indicated that Decision Tree had better F1 score; hence Decision Tree model was better than KNN. Decision Tree predicted the type of cancer better in head and neck regions.

It can be re-written in a compact representation:
Model ($_{\text{HNC type}}$) = M($_{\text{Decision Tree}}$) > M_{KNN}

If F1 Score > Accuracy in a model THEN the model is good. The FI score of Decision Tree was greater than F1 Score of KNN, hence between Decision Tree and KNN models, Decision Tree performed better than k-NN in classification of head and neck cancer types.

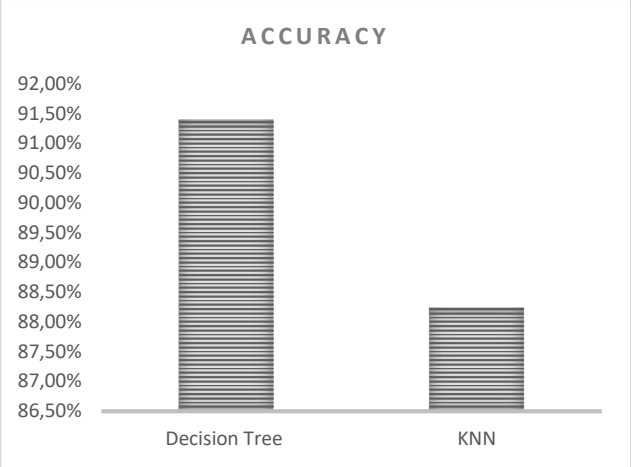

Fig. 2. Accuracy of the models (Decision Tree and K-NN)

VI. Conclusion

Early presentation and commencement of treatment options in cancer patients is a strong factor of good prognostic outcome.

The paper observed/identified features that are important in the diagnosis of head and neck cancer types. This will help the paramedics being the first point of contact in some remote locations to identify the features and refer.

The integration of machine learning algorithm for diagnosis of HNC in a primary health centres will invariably shorten the referral time with ultimate prompt referral to the ENT/Head and Neck specialists in the tertiary health institutions. Overall, the rate of morbidity and mortality of patients with HNC due to late presentation would be stemmed.

Acknowledgment

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

References

[1] Altman, N. S. (1992). An Introduction to kernel and kernel-Neighbour non parametric regression. The American Statistician 46(3): 175-185. Doi:10.1080/00031305.1992.10475879

[2] Baatenburg, D R. J., Hermans, J., Molenar, J., Briaire, J. J. and Le Cessie, S. (2001). Prediction of Survival in Patients with Head and Neck Cancer. National Center for Biotechnology Information, John Wiley & Sons, Inc.

[3] Benjamín, M. and Carlos, H. M. (2013). Prediction System of Larynx Cancer. *The Fourth International Conference on Computational Logics, Algebras, Programming, Tools, and Benchmarking*

[4] Durairaj, M. and Deepika R. (2015). Prediction of Acute Myeloid Leukemia Cancer using Data Mining- A Survey. *International Journal of Emerging Technology and Innovative Engineering* 1(2)

[5] Erinoso, O. A., Okoturo, E., Gbotolorun, O. M., Effiom, O. A., Awolola, N. A., Soyemi, S. S. and Oluwakuyide, R. T. (2016). Emerging Trends in the Epidemiological Pattern of Head and Neck Cancers in Lagos, Nigeria. Annals Medical Health Science Res. 6(5): 301–307

[6] *GBD (2015). Mortality and Causes of Death, Collaborators.* Global, Regional, and National Life Expectancy, All-Cause Mortality, and Cause-Specific Mortality for 249 Causes of Death, 1980-2015: A Systematic Analysis for the Global Burden of Disease Study 2015. *Lancet. 388 (10053): 1459–1544.*

[7] Hagedoorn, M. and Molleman, E. (2006). Facial Disfigurement in Patients with Head and Neck Cancer: The Role of Social Self-Efficacy. *American Psychological Association.* 25(5): 643–647

[8] Hussein, A., Shigeo, K. and Yasuhiro, A. (2002). Development and Applications of Decision Trees. Expert Systems. 1(1):53-77

[9] Jacqueline, A. E., L., Jos T., Martine, K., Patricia, D., Mark, H.H., Kramer, P. M., and Weijs, and Leemans, C. R. (2015). Prediction Model to Predict Critical Weight Loss in Patients with Head and Neck Cancer during (Chemo) Radiotherapy. Oral Oncology, 52(1): 91-96

[10] *Jiawei H., Micheline, K. and Jian, P. (2011). Data Mining: Concepts and Techniques 3rd Edition*

[11] Han, J., Kamber, M. and Pei, J. (2012). Data Mining: Concepts and Techniques, 3rd Edition. Elsevier, Amsterdam

[12] John, C. W., Mark, N. G., and Janet A. W. (2000). Stell and Maran's Head and Neck Surgery. Butterworth Heinemann. Fourth Edition

[13] Opubo, B. D., Abayomi, O. S. and Wasiu, L. A. (2009). Current Evidence on the Burden of Head and Neck Cancers in Nigeria. *Head Neck Oncol*.

[14] Rajeswari, B. and Aruchamy, R. (2014). Survey on Data Mining Algorithms to predict Leukemia Types *International Journal for Research Science Engineering and Technology.* (IJRSET) 2(5): 42-46

[15] Sami, P. M., John, S. S., Tareck, A., Louis, G., Eric, B., Olguta, E. G., Denis, S., Louise, L., Edith, F., Phuc, F. N., and Apostolos, C. (2015). Predicting Depression and Quality of Life among Long-Term Head and Neck Cancer Survivors. *American Academy of Otolaryngology*—Head and Neck Surgery. 152(1): 91–97

[16] *Shakhnarovish, D. and Indyk (2005). Nearest-Neighbor Methods in Learning and Vision.*

O. Olabode is a Professor of Computer Science from the Federal University of Technology, Akure. He holds a B.Tech degree in Industrial Mathematics and M.Tech. Degree in Computer Science in 1991 and 1999 respectively from The Federal University of Technology, Akure, Nigeria. In 2005, he obtained a PhD degree in Computer Science and in 2015 he obtained an M.Tech degree in Statistics both from the Federal University of Technology, Akure, Nigeria.

He is a member of some professional bodies such as the Computer Professionals (Registration Council of Nigeria), Computer Society of Nigeria among others. He is a reviewer of some academic and professional journals such as International Journal of Computer Systems and Applications, British Journal of Mathematics & Computer Science, Issues in Business Management and Economics. Prof. Olabode has both local and international teaching and research experience. His research area includes eCommerce, Machine learning and Softcomputing.

Adebayo Adetunmbi is a Professor in the Department of Computer Science, Federal University of Technology, Akure (FUTA) where he obtained his PhD in Computer Science in 2008. He was a recipient of CAS-TWAS postgraduate fellowship at the Institute of Computing Technology, Beijing in 2006 and a Visiting Scholar to Massachusetts Institute of Technology in 2012. He has over one hundred publications in reputable peer-reviewed journals and conference proceedings. His research interests are Data Science, Computational linguistics and Information security. He is a member of IEEE computer society and Computer Professional Council of Nigeria

Folake Akinbohun works with Rufus Giwa Polytechnic, Owo, Ondo State, Nigeria. She bags Master of Technology in Computer Science from Federal University of Nigeria in 2012. She is pursuing 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 (AI).

Dr. Ambrose Akinbohun 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.