International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

Comparative Analysis of the C4.5 and ID3 Decision Tree Algorithms for Disease Symptom Classification and Diagnosis

Nicodemus Nzoka Maingi[1], Ismail Ateya Lukandu[1], Matilu Mwau[2]
[1]Faculty of Information Technology, Strathmore University, Nairobi, Kenya
[2]Kenya Medical Research Institute (KEMRI), Ministry of Health, Nairobi, Kenya

Abstract— *Most disease surveillance outfits and authrorities around the world battle with one key challenge – the useful and objective handling and processing of the huge sets of disease data being generated on a regular basis as their personnel exercise their disease surveillance mandate. Many theories have been put forth on how best this could be tackled. Among these, is the use of information technology and mathematical theories and concepts to alleviate the problem. One of the most solid and promising methods includes the use of artificial intelligence techniques to help break down and make good sense of the data sets. This research looks to compare the usage of the C4.5 and the ID3 decision tree theory concepts as means of tackling making the best of disease surveillance data. The C4.5 and ID3 algorithms provide a method of breaking down the data and generating (among other useful information) the entropies and information gains of some predefined variables from huge sets of disease outbreak data. Once the information gain scores for the variables are computed, they can be easily ranked to determine the variable to define the root node in the decision tree, as the rest of the variables follow through as leaf nodes. Notably, there will be two sets of entropies and information gains; one from the C4.5 algortihm and the other from the ID3 algorithm. Both decision trees shall have validation steps after each branch pass to determine whether it is time to stop growing it or not. This is one of the mechanisms employed here to avoid overfiting of the decision tree (especially for the ID3 algorithm).*

Keywords— *C4.5, ID3, Decision Tree, Disease Symptom Burden Variables, Entropy, Information Gain, Notifiable Disease, Overfitting.*

I. INTRODUCTION

The Disease Surveillance and Response Unit (DSRU) falls under the Ministry of Health in Kenya and is mandated to spearhead efforts in the management of diseases outbreaks within a dynamic target disease list, popularly referred to as a notifiable disease list for Kenya. The DSRU works hand-in-hand with various local, regional and global partners in its efforts to combat notifiable disease outbreaks. Artificial intelligence concepts have been used and applied successfully in many fields. The choice of using decision trees is driven by the modular approach to this research; the breaking down of each notifiable disease into its requisite symptom burden variables such that disease diagnosis will be done by mapping together a predefined cluster of symptom burdens variables. As per [10], surveillance data are collected at the health facility – the first level of contact of the patient with the health system – then analyzed, interpreted, and used to inform action.

Information and communication technology (ICT) has improved, a development that presents new opportunities and possibilities for better disease surveillance data processing and visualization. [8] presents automated subsystems that take diagnosis a notch higher; better disease surveillance efforts can improve the livelihoods and life expectancies of a people far beyond traditional practice. With the age of industrialization, the modern industrialized workspace has erased borders and made the movement and spread of people much more rapid; the same applying to the spreading of any ailments or bugs that the people carry. [2] looks at big data as the next gold mine; providing endless opportunities for the crunching of data into meaningful, useful and actionable knowledge. Applied artificial intelligence and machine learning can give this fight against disease outbreaks the positive impetus it requires.

This research study also seeks to answer the research questions:

How do the C4.5 and ID3 decision tree algorithms compare in the diagnostics and classification of outbreak diseases?

Which is of the two (C4.5 and ID3) algorithms provides the best basis for the symptom burden classification?

II. LITERATURE REVIEW

A. Machine Learning and Decision Trees Theory and Practice

Artificial intelligence and by extension, machine learning have been used as game changers in the handling of huge, complex and constantly growing data sets across many disciplines. The use of machine learning and crowd-sourcing techniques have defined a new panorama through which disease outbreak data speaks volumes [5]. The collation and processing of such data sets can drive the early outbreak detection whilst increasing public awareness of disease outbreaks. When artificial intelligence techniques are applied, then the data sets create various useful perspectives, making it easier to undertake useful and far-reaching interventions [2].

There are different ideas around the utilization of different clinical disease parameters to inform the construction of decision trees; this is thought to enhance the advancement of intuitive diagnostic algorithms capable of comfortably managing disease outbreak data [9]. This push could lead to the design and development of an automated diagnostics system [7]. Even better, the panorama is even more enriched if there is a comparative analysis of both the C4.5 and the ID3

Nicodemus Nzoka Maingi, Ismail Ateya Lukandu, and Matilu Mwau, "Comparative Analysis of the C4.5 and ID3 Decision Tree Algorithms for Disease Symptom Classification and Diagnosis," *International Research Journal of Advanced Engineering and Science*, Volume 4, Issue 2, pp. 188-194, 2019

International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

algorithmic perspectives. Both these algorithms could help define a guiding set of structured questions on the clinical, physical and historical conditions of a particular disease on a patient in order to assist in better diagnostics and management of such disease outbreaks.

B. Web- and Mobile-Based Technology

Automation of disease surveillance efforts is a critical activity in the fight against disease outbreaks. This can be achieved through the use of both web- and mobile-based tools. [2] portrays the use of web-based electronic information sources as a critical driver in the early event detection and support situational awareness by providing highly localized and current information about outbreaks. This can only be useful once the disease parameters to capture and monitor are established. Any new information and/or indicators can be continually and dynamically updated onto the web- or mobile-based platform to reflect reality of the current designations.

III. METHODOLOGY

The methodology used in the research is mainly experimental research, boosted by evolutionary prototyping and modeling. Experimental research analysis is one of the branches of quantitative research methods; working with different data variables ([2]; [3]; [4]). It mainly points to the systematic, theoretical analysis of the methods applied to a field of study [6]. Some artificial intelligence techniques such as machine learning and decision trees theory have also been employed in the research.

Both the C4.5 and ID3 algorithms for calculating entropy and information gains have been applied. Once the entropy and the information gains (due to the various disease burdens variables) are determined, the decision trees are constructed. The nodes, branches and leaves indicate the variables, conditions, and outcomes, respectively [1]. In both cases, the attribute with the smallest entropy or the largest information gain is placed as the root node(s), then the rest follow on as the leaf nodes in order.

The research follows the process flow outlined below:

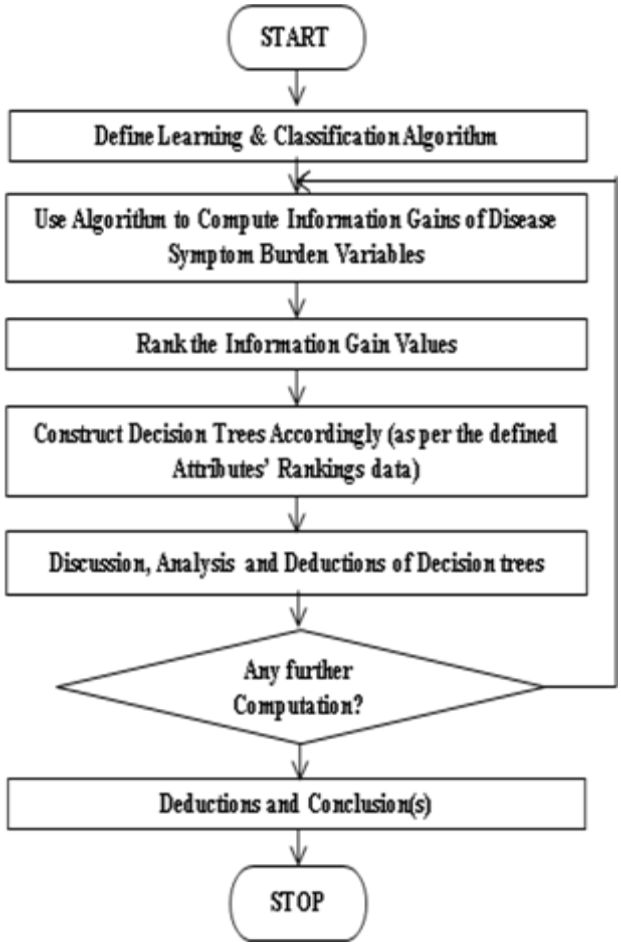

Fig. 1.

In the end, the decision in the entropy and information gains define the disease classification of the disease variables e.g. Measles, Dengue Fever etc.

Entropy determination:

$$\text{Entropy (Decision)} = \sum_{n=1}^{\infty} -p(\text{Decision}).\log_2 p(\text{Decision}) \qquad \text{Equation (1)}$$

Information gains determination:

$$\text{I G (Decision, Variable)} = \text{Entropy(Decision)} - \sum_{n=1}^{\infty} [p(\text{Decision}|\text{Variable}).\text{Entropy(Decision}|\text{Variable})] \qquad \text{Equation (2)}$$

Nicodemus Nzoka Maingi, Ismail Ateya Lukandu, and Matilu Mwau, "Disease Outbreak Classifier and Prioritization Algorithm using the C4.5 Machine Learning Decision Trees," *International Research Journal of Advanced Engineering and Science*, Volume 4, Issue 2, pp. 188-194, 2019.

International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

TABLE 1. Nairobi Country Disease Symptom Burdens Variable Nominal Data (2015 – 2018)

	Overall Aggregated Disease Data							
	Symptomatic Observation Code Values							
Disease Symptom Variables	B	G	M	N	O	P	R	S
Control Case Zero	0	0	0	0	0	0	0	0
Adverse Effects Following Immunization	0	0	0	0	0	0	0	0
Anthrax	28	28	0	0	28	28	28	0
Cholera	188	188	0	0	188	0	0	0
Dengue Fever	16	16	0	0	4	16	16	2
Dysentery	1,028	1,028	0	0	0	1,028	0	0
Guinea Worm	10	10	0	0	0	0	0	10
Measles	138	138	0	138	133	138	138	138
Neonatal Tetanus	3	3	3	0	0	0	0	0
Plague	10	10	0	0	10	10	10	10
Rift Valley Fever	7	7	0	7	7	7	0	7
Severe Acure Respiratory Infection	10	0	0	0	0	0	10	9
Viral Haemorrahic Fever	15	15	0	0	15	15	15	15
Yellow Fever	45	45	0	0	45	45	0	45
Polio	11	11	0	0	0	8	0	0
Acute Jaundice	86	86	0	0	0	86	0	0
Acute Malnutrition	798	798	798	0	0	0	798	798
Malaria	138,978	138,978	0	0	0	133,246	0	138,978
Meningitis	22	22	0	0	0	22	0	0
Rabies	23	12	23	0	0	23	0	0
Tuberculosis	1,363	1,363	0	0	0	0	1,363	345
Typhoid	33,715	33,715	0	0	0	33,715	33,715	0
OTHERS	13,980	13,980	13,980	13,980	13,980	13,980	13,980	13,980

TABLE 2. Nairobi County C4.5 Entropy, Information Gains and Rankings

Disease Symptom Variables	B	G	M	N	O	P	R	S
Entropy (Decision)	4.5236	4.5236	4.5236	4.5236	4.5236	4.5236	4.5236	4.5236
Information Gains (Decision/Variable)	4.3496	4.4366	1.0144	0.7654	2.2060	3.4801	2.3451	2.8691
Information Gains Rankings	2	1	7	8	6	3	5	4

TABLE 3. Nairobi County Disease Symptom Burden Variables Rankings

Information Gain Rankings	Disease Burden Variables
1	G
2	B
3	P
4	S
5	R
6	O
7	M
8	N

TABLE 4. Legend of Disease Burdens Variables

B	Bodily Disease Manifestations
G	Gastrointestinal Disease Manifestations
M	Muscular Disease Manifestations
N	Nasal Disease Manifestations
O	OTHER Disease Manifestations
P	Pain Disease Manifestations
R	Respiratory Disease Manifestations
S	Skin Disease Manifestations

Nairobi County C4.5 Decision Tree Construction (based on Information Gains Rankings of Disease Burden Variables)

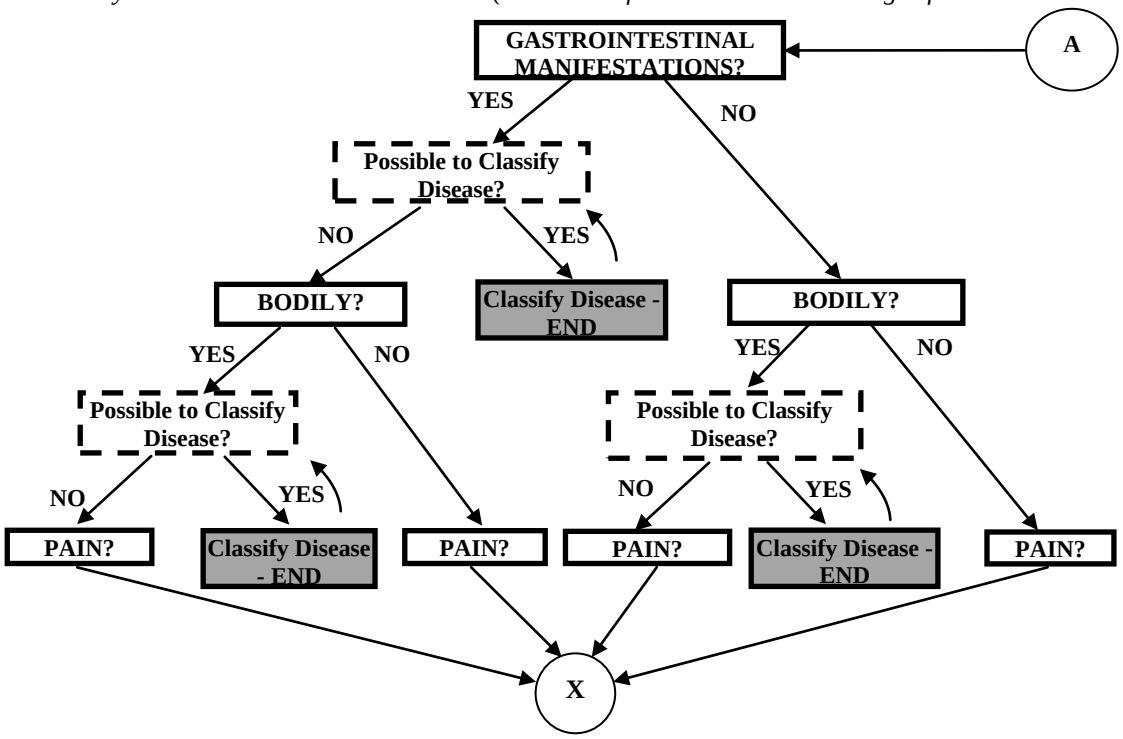

Fig. 2. C4.5 Decision Tree

Nicodemus Nzoka Maingi, Ismail Ateya Lukandu, and Matilu Mwau, "Disease Outbreak Classifier and Prioritization Algorithm using the C4.5 Machine Learning Decision Trees," *International Research Journal of Advanced Engineering and Science*, Volume 4, Issue 2, pp. 188-194, 2019.

International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

Nairobi County C4.5 Decision Tree Construction (Continuation)

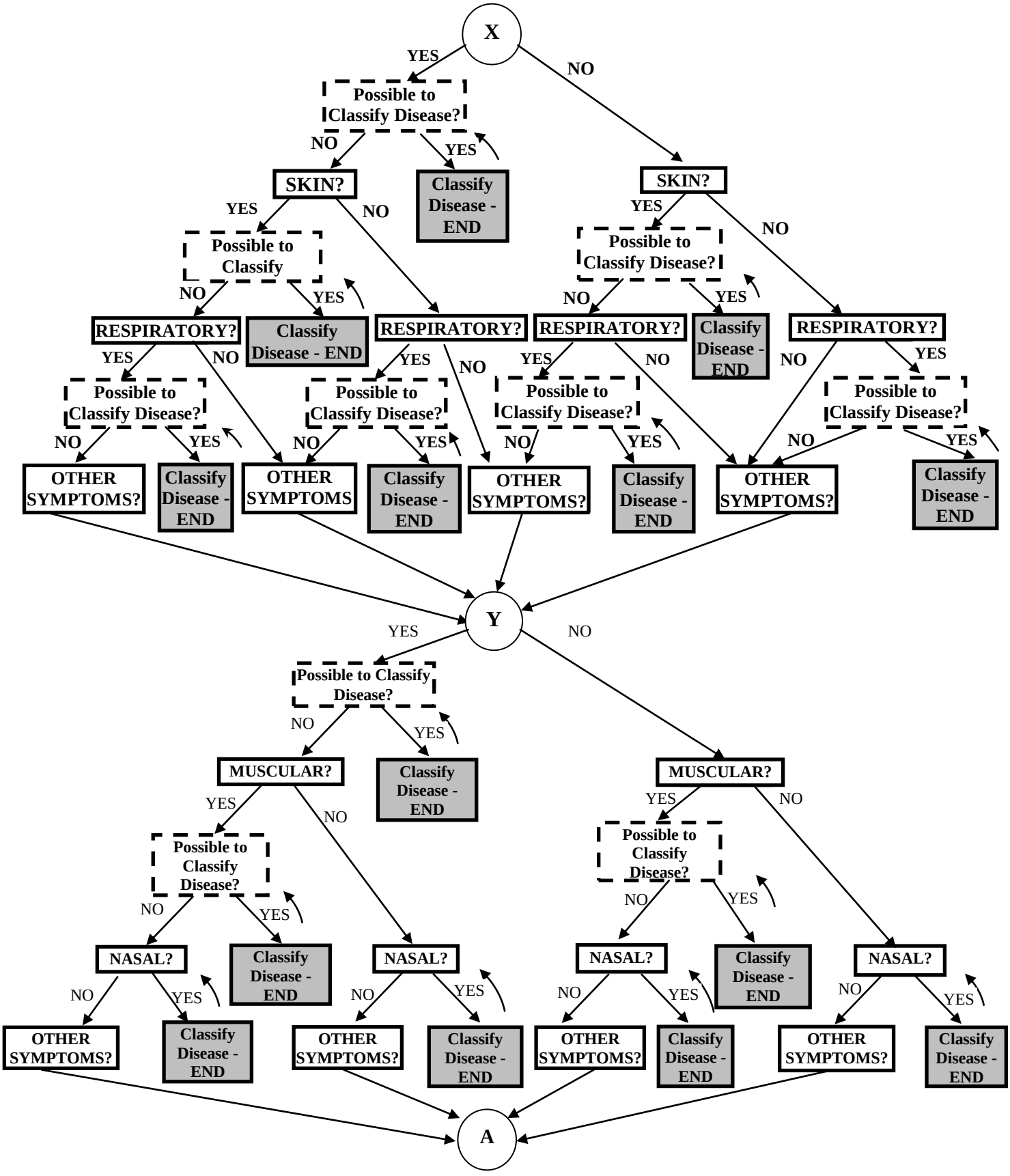

Fig. 3. C4.5 Decision Tree (Continuation)

Nicodemus Nzoka Maingi, Ismail Ateya Lukandu, and Matilu Mwau, "Disease Outbreak Classifier and Prioritization Algorithm using the C4.5 Machine Learning Decision Trees," *International Research Journal of Advanced Engineering and Science*, Volume 4, Issue 2, pp. 188-194, 2019.

International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

TABLE 5. Nairobi Country Disease Symptom Burdens Variable Qualitative Data (2015 – 2018)

	Overall Aggregated Symptomatic Observation Code Values							
Disease codes	B	G	M	N	O	P	R	S
Control Case Zero	NONE	NONE	NONE	NONE	NONE	NONE	NONE	NONE
Adverse Effects Following Immunization	NONE	NONE	NONE	NONE	NONE	NONE	NONE	NONE
Anthrax	HIGH	HIGH	NONE	NONE	HIGH	HIGH	HIGH	NONE
Cholera	VERY HIGH	VERY HIGH	NONE	NONE	VERY HIGH	NONE	NONE	NONE
Dengue Fever	HIGH	HIGH	NONE	NONE	LOW	HIGH	HIGH	LOW
Dysentery	VERY HIGH	VERY HIGH	NONE	NONE	NONE	VERY	NONE	NONE
Guinea Worm	MEDIUM	MEDIUM	NONE	NONE	NONE	NONE	NONE	MEDIUM
Measles	VERY HIGH	VERY HIGH	NONE	VERY	VERY HIGH	VERY	VERY	VERY
Neonatal Tetanus	LOW	LOW	LOW	NONE	NONE	NONE	NONE	NONE
Plague	HIGH	HIGH	NONE	NONE	HIGH	HIGH	HIGH	HIGH
Rift Valley Fever	MEDIUM	MEDIUM	NONE	MEDIUM	MEDIUM	MEDIUM	NONE	MEDIUM
Severe Acute Respiratory Illness	HIGH	LOW	NONE	NONE	NONE	NONE	HIGH	MEDIUM
Viral Hemorrhagic Fever	HIGH	HIGH	NONE	NONE	HIGH	HIGH	HIGH	HIGH
Yellow Fever	HIGH	HIGH	NONE	NONE	HIGH	HIGH	NONE	HIGH
Polio	HIGH	HIGH	NONE	NONE	NONE	MEDIUM	NONE	NONE
Acute Jaundice	HIGH	HIGH	NONE	NONE	NONE	HIGH	NONE	NONE
Acute Malnutrition	VERY HIGH	VERY HIGH	VERY	NONE	NONE	NONE	VERY	VERY
Malaria	VERY HIGH	VERY HIGH	NONE	NONE	NONE	VERY	NONE	VERY
Meningitis	HIGH	HIGH	NONE	NONE	NONE	HIGH	NONE	NONE
Rabies	HIGH	HIGH	HIGH	NONE	NONE	HIGH	NONE	NONE
Tuberculosis	VERY HIGH	VERY HIGH	NONE	NONE	NONE	NONE	VERY	VERY
Typhoid	VERY HIGH	VERY HIGH	NONE	NONE	NONE	VERY	VERY	NONE
OTHERS	VERY HIGH	VERY HIGH	VERY	VERY	VERY HIGH	VERY	VERY	VERY

TABLE 6. ID3 Entropy, Information Gains and Rankings

Disease symptom Variables	B	G	M	N	O	P	R	S
Entropy (Decision)	4.5236	4.5236	4.5236	4.5236	4.5236	4.5236	4.5236	4.5236
Information Gain (Decision|Variable)	1.8619	1.9788	0.9274	0.6784	1.6515	1.8449	1.4225	1.9508
Information Gain Rankings	3	1	7	8	5	4	6	2

TABLE 7. Disease Symptom Burden Variables Rankings

Rankings	Disease Burden Variables
1	G
2	S
3	B
4	P
5	O
6	R
7	M
8	N

Nicodemus Nzoka Maingi, Ismail Ateya Lukandu, and Matilu Mwau, "Disease Outbreak Classifier and Prioritization Algorithm using the C4.5 Machine Learning Decision Trees," *International Research Journal of Advanced Engineering and Science*, Volume 4, Issue 2, pp. 188-194, 2019.

International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

Nairobi County ID3 Decision Tree Construction (based on Information Gains Rankings of Disease Burden Variables)

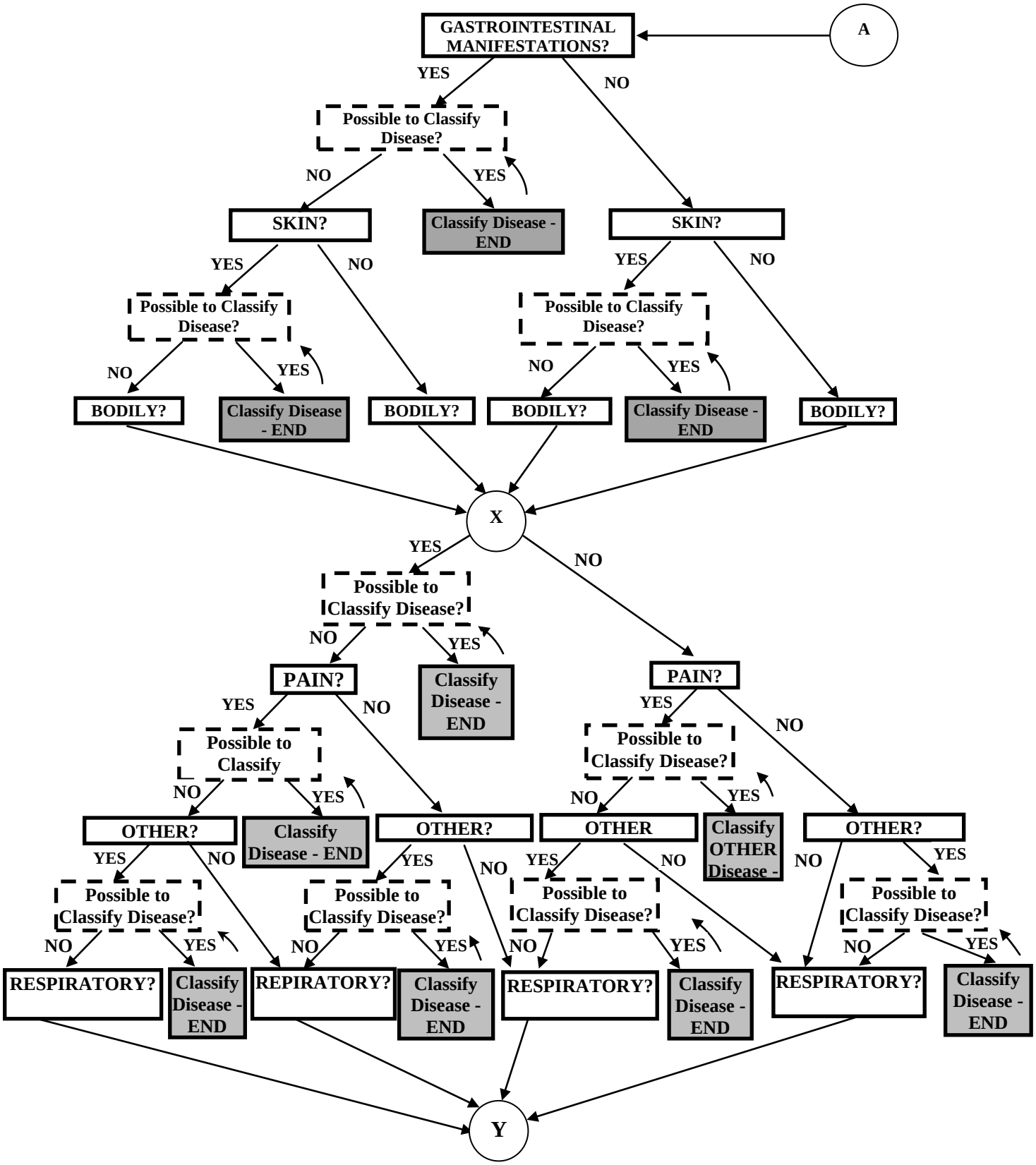

Fig. 4. ID3 Decision Tree

Nicodemus Nzoka Maingi, Ismail Ateya Lukandu, and Matilu Mwau, "Disease Outbreak Classifier and Prioritization Algorithm using the C4.5 Machine Learning Decision Trees," *International Research Journal of Advanced Engineering and Science*, Volume 4, Issue 2, pp. 188-194, 2019.

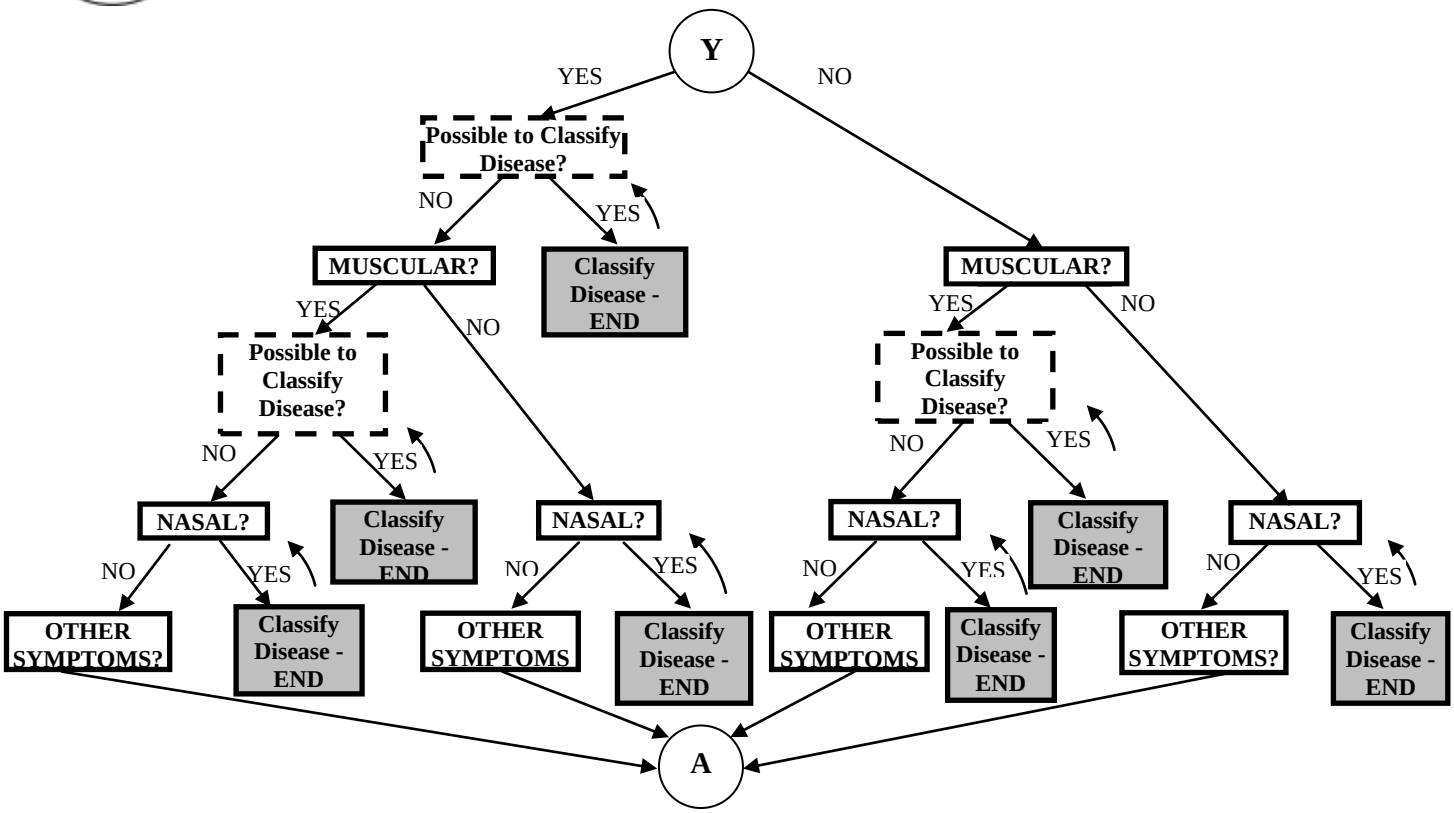

Fig. 5. ID3 Decision Tree (Continuation)

IV. Results and Discussion

From the results, the two decision trees closely compare, with the exception of some slight variations in ranking of some of the disease variables. Either of the two decision trees can be used to classify diseases. This gives the disease surveillance efforts alternatives to disease classification. This can hopefully give the fight against disease outbreaks the impetus it needs since the model defined here can help make better and informed ways of usefully aggregating the disease data and packaging it in a manner that helps the medical personnel make the most of out of it.

Finally, the two research questions are answered: how the C4.5 and ID3 decision tree algorithms can easily be compared. Additionally, none of the two can be said to be better than the other; each of the two defines its own defined decision tree to classify diseases.

V. Conclusion

Comparative decision trees can be constructed and used to define disease classifications. This can be used to drive alternative plans and policies to help improve healthcare delivery especially with regard to the management of disease outbreaks by driving better disease research for the development of better medicines as well as pushing towards the elimination of some of the diseases. Finally, the model drives us to conclude null hypothesis (that *Disease Symptom Burden variables can be mainly used to determine the information gains and consequent ranking for decision tree nodes for disease classification and prediction)* holds.

References

[1] Agrawal, G. L., & Gupta, H. (2013). Optimization of C4. 5 decision tree algorithm for data mining application. International Journal of Emerging Technology and Advanced Engineering, 3(3), 341-345.

[2] Brownstein, J. S., Freifeld, C. C., Reis, B. Y., & Mandl, K. D. (2008). Surveillance Sans Frontieres: Internet-based emerging infectious disease intelligence and the HealthMap project. PLoS medicine, 5(7), e151.

[3] Creswell, J. W. (2013). Research Design: Qualitative, Quantitative, and Mixed Methods Approaches. Sage Publications.

[4] Harland, D. J. (2011). Science, Technology, Engineering and Mathematics (STEM) Student Research Handbook. NSTA Press.

[5] Hay, S. I., George, D. B., Moyes, C. L., & Brownstein, J. S. (2013). Big data opportunities for global infectious disease surveillance. PLoS medicine, 10(4), e1001413.

[6] Howell, K. E. (2012). An Introduction to the Philosophy of Methodology. Sage.

[7] Rahman, R. M., & Hasan, F. R. M. (2011). Using and comparing different decision tree classification techniques for mining ICDDR, B Hospital Surveillance data. Expert Systems with Applications, 38(9), 11421-11436.

[8] Roser, M. (2015). Life expectancy. Our World in Data. Accessed on 12th July 2017 from http://ourworldindata.org/data/population-growth-vital-statistics/life-expectancy.

[9] Tanner, L., Schreiber, M., Low, J. G., Ong, A., Tolfvenstam, T., Lai, Y. L., & Simmons, C. P. (2008). Decision tree algorithms predict the diagnosis and outcome of dengue fever in the early phase of illness. PLoS neglected tropical diseases, 2(3), e196.

[10] World Health Organization (WHO). (2015). Report Global Surveillance of Epidemic-prone Infectious Diseases (WRGSEID) Department of Communicable Disease Surveillance and Response