INTERNATIONAL JOURNAL OF
INNOVATIVE RESEARCH AND KNOWLEDGE
ISSN-2213-1356 www.ijirk.com

Disease Outbreak Classifier and Prioritization Algorithm using the C4.5 Machine Learning Decision Trees

Nicodemus Nzoka Maingi
Lecturer, Faculty of Information Technology, Strathmore University, Nairobi, Kenya

Ismail Ateya Lukandu
Senior Lecturer and Dean of Research and Innovation, Faculty of Information Technology,
Strathmore University, Nairobi, Kenya

Matilu Mwau
Senior Research Scientist, Kenya Medical Research Institute (KEMRI)
Nairobi, Kenya

Abstract

Disease outbreak data sets can sometimes present an overwhelming challenge to the concerned authorities. This mainly driven by among other factors, that disease outbreaks are mostly random, exhibiting very unpredictable and sometimes irrepeatable patterns in addition to the volumes of data they generate. But at the core of these reported diseases, are symptom burdens that characterize each disease; each diease can be broken down into its unique, and irrepeatable constituent symptoms it manifests. These symptoms are now dubbed symptom burden variables and they can then be used as atomic representatives of a disease. By breaking each disease to its particular symptomatic burdens, it is possible to use the broken down data sets to define a predictive model or algorithm to enable authorities know and plan better on what diseases to anticipate and prioritize. Defined here is a learning and classification algorithmic model that relies on the different diseases' symptom, using these to enable the computation of each disease burdens' variable's information gains or ratios and their consequent ranks to enable the dynamic construction of relevant and corresponding decision trees. The ranked variables act as decision tree nodes (both parent and child), making it possible to define objective decision trees that gives a clearer hierachy of each symptom burden (from the root node [most critical variable], down to the leaf nodes [least critical variables]. Using this algorithm, it is possible to (over a period of time) keep generating dynamic and up-to-date decision trees as the outbreak situation unfolds over time to inform disease outbreak mitigating efforts.

Keywords: *Decision Trees, Disease Symptom Burden Variables, Entropy, Information Gain.*

1. Introduction

The Disease Surveillance and Response Unit (DSRU) in Kenya is mandated to manage disease outbreaks for a set of target diseases, usually collectively referred to as the notifiable disease list; notifiable derived from the fact that any one (1) to five (5) occurrences of any of these disease must be reported to the health authorities and samples taken from patients for testing and confirmation at the national laboratory. Kenya's notifiable disease list has a set of 14 diseases (see Table 4).

Roser (2015) avers that life expectancy has increased rapidly since the age of enlightenment. This has partly been as a result of better nutrition coupled with better healthcare practice, especially the management of disease outbreaks. In all these efforts, artificial intelligence has played its part in improving the efficiency of healthcare service provision. Investigators have created programs that simulate expert human reasoning; this in a bid to overcome limitations intrinsic in the traditional computer-aided diagnosis (Roser, 2015). Still, disease outbreak surveillance remains one of the most critical aspects of disease epidemics management in the modern world. Whilst the traditional wait-and-see and reactionary approach worked in the past, disease strains and their frequency have now become more complex and more prevalent, prompting for better methods of mitigation. Additionally, the modern industrialized workplace has caused more cross-border movement of people, in the exchange of goods and services, and with them, all manner of disease strains, driving the demand for more proactive approaches. Different nation's disease surveillance authorities are constantly amassing huge repositories of disease outbreak data. These volumes require some very creative approaches if only to maximize on the value of the huge data sets. The next decade of disease surveillance research will require new approaches to effectively make sense of rapid, massive quantities of complex, and multi-dimensional data sets (Neill, 2012). Some of these approaches include the use of artificial intelligence and machine learning concepts to model and predict possible disease outbreak patterns well in advance.

2. Problem Statement

In most nations, the efforts to combat disease outbreaks faces a myriad of challenges, ranging from limited or no funding, to poor or lack of requisite training and knowledge, to the development of new and mutated disease strains, among other challenges. This research looks to address one of the challenges; how to deal with the huge sets of disease data sets that are generated over time as the health personnel expend their efforts in fighting and managing disease outbreaks.

3. Theoretical Framework

Machine learning provides one of the most efficient mechanisms to enable the handling and/or processing the huge sets of disease data to influence meaningful deductions and interventions in the disease surveillance space of any nation. According to Hay (2013), advances in machine learning and the use of crowd sourcing have opened up the possibility of the development of a dynamic atlas of infectious diseases. In the end, the use of other nonconventional sources of disease surveillance data can enhance the early outbreak detection as well as increasing public awareness of disease outbreaks to complement the application of artificial intelligence techniques to maximize visibility and inform useful interventions (Brownstein, 2008).

Tanner (2008) argues for the use of different clinical disease parameters to analyze and model disease data as a useful driver in the development of intuitive diagnostic algorithms for handling disease outbreak data. This algorithm could consequently inform the development of an intuitive computer application that could be useful in the diagnosis and determination of how critical an outbreak is and what symptom burdens could be of critical importance to lay focus upon in mitigating efforts; diagnosis could be by making use of up-to-date but historical data kept in the surveillance database (Rahman, 2011).

4. Conceptual Framework

Decision trees provide a repeatable and objective mechanism of arriving at critical decisions involving several complexly linked variables. According to Patel (2015), decision tree learning algorithms have been successfully used in expert systems in capturing knowledge, with most decision tree classifiers being designed for the classification of cases with categorical or boolean class data sets.

There will be use of entropies and information gains of disease symptom burden variables. Entropy defines the amount of uncertainty or disorder in a given system (Russel, 2009).

This research study also seeks to answer the research question: *What is the order of disease burden variables in determining what disease or diseases to prioritize in fighting disease outbreaks?*

Null Hypothesis, H_o: *the Disease Symptom Burden variables can be used to determine the information gains and consequent rankings for decision tree nodes for disease classification and prediction.*

The algorithm defined here could drive the design and construction of a priority list of symptoms (of interest) to be used as the basis for diagnosis at a country's ports of entry or exit.

The priority list of symptoms could be key to the automation of the disease surveillance operations. This can be achieved through the use of web- and mobile-based tools. Brownstein (2008) portrays the use of web-based electronic information sources as a critical driver in the early event detection and support situational awareness by providing current, highly local information about outbreaks. This can only be useful once the disease parameters or variables 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 on the ground.

5. Methodology

The methodology used in the research is mainly experimental research, boosted by prototyping and modeling. Experimental research analysis is one of the branches of quantitative research methods; working with different data variables (Brownstein, 2008; Harland, 2011; Creswell, 2013). It mainly points to the systematic, theoretical analysis of the methods applied to a field of study (Howell, 2012). Some artificial intelligence techniques such as machine learning and decision trees theory have also been employed in the research.

Artificial intelligence and decision tree concepts, mainly the C4.5 algorithm for defining entropy and information gains has been applied. Once the entropy and the information gains (due to the various disease burdens variables) are determined, the decision trees are constructed based on the symptom burden variables as the tree nodes – starting with the variables with the highest information gain down to those with the lowest. The nodes, branches and leaves indicate the variables, conditions, and outcomes, respectively (Agrawal, 2013). Generally, 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, with the tree branching in boolean fashion until either all variable are exhausted or a disease can be classified without going further down the tree.

The research follows the process flow outlines below:
In this research, the decision is used to define the determined disease classification e.g. Measles, Dengue Fever etc based on each of its symptom burden variables.

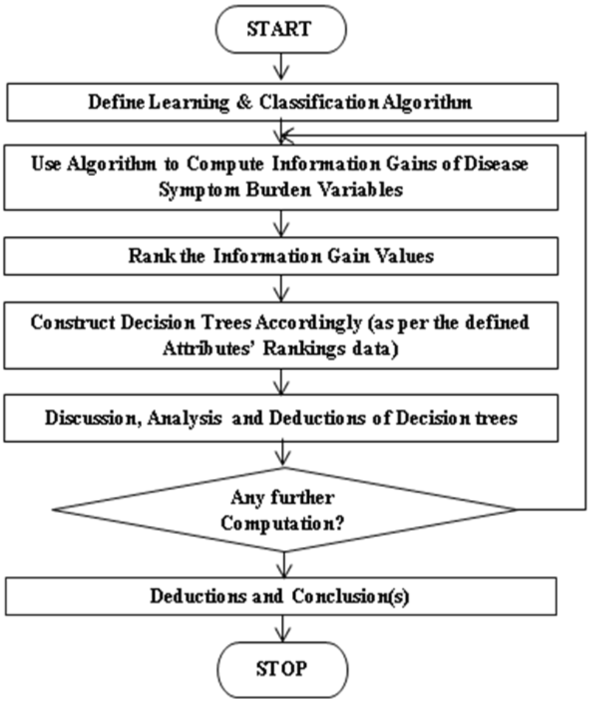

START
Define Learning & Classification Algorithm
Use Algorithm to Compute Information Gains of Disease Symptom Burden Variables
Rank the Information Gain Values
Construct Decision Trees Accordingly (as per the defined Attributes' Rankings data)
Discussion, Analysis and Deductions of Decision trees
Any further Computation?
Deductions and Conclusion(s)
STOP

Figure 1: Algorithm and decision tree activity process flow

The formulae used in the research are defined below:

Entropy determination:

Equation (1)

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

Russel (2009)

Information Gains determination:

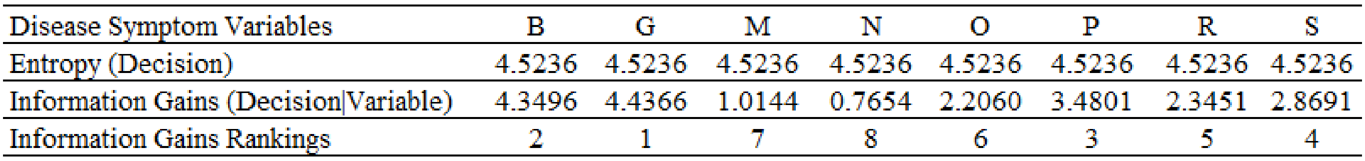
Equation (2)

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

Russel (2009)

6. Research Results

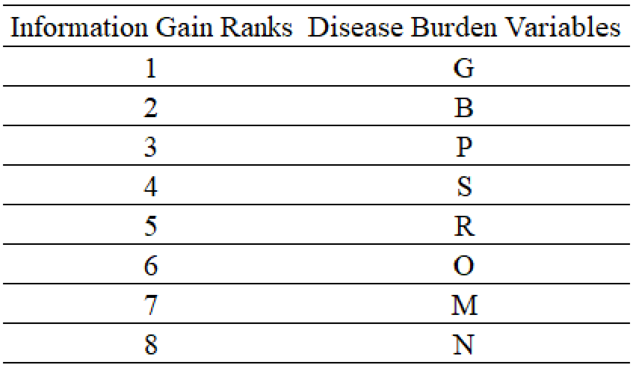
Table 1: Entropy, Information Gains and Rankings



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

Source: Calculations from R and Microsoft Excel based on DHIS 2015 – 2018 Raw Data (Table 1).

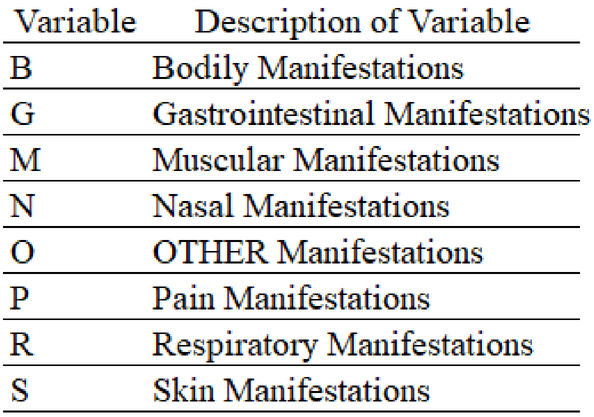
Table 2: Disease Symptom Burden Variables Rankings



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

Source: Calculations from R and Microsoft Excel based on DHIS 2015 – 2018 Raw Data (Table 1).

Table 3: Legend of Disease Burdens Variables



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

Source: Mayo Clinic Online Database of Diseases and their Symptoms (https://www.mayoclinic.org/)

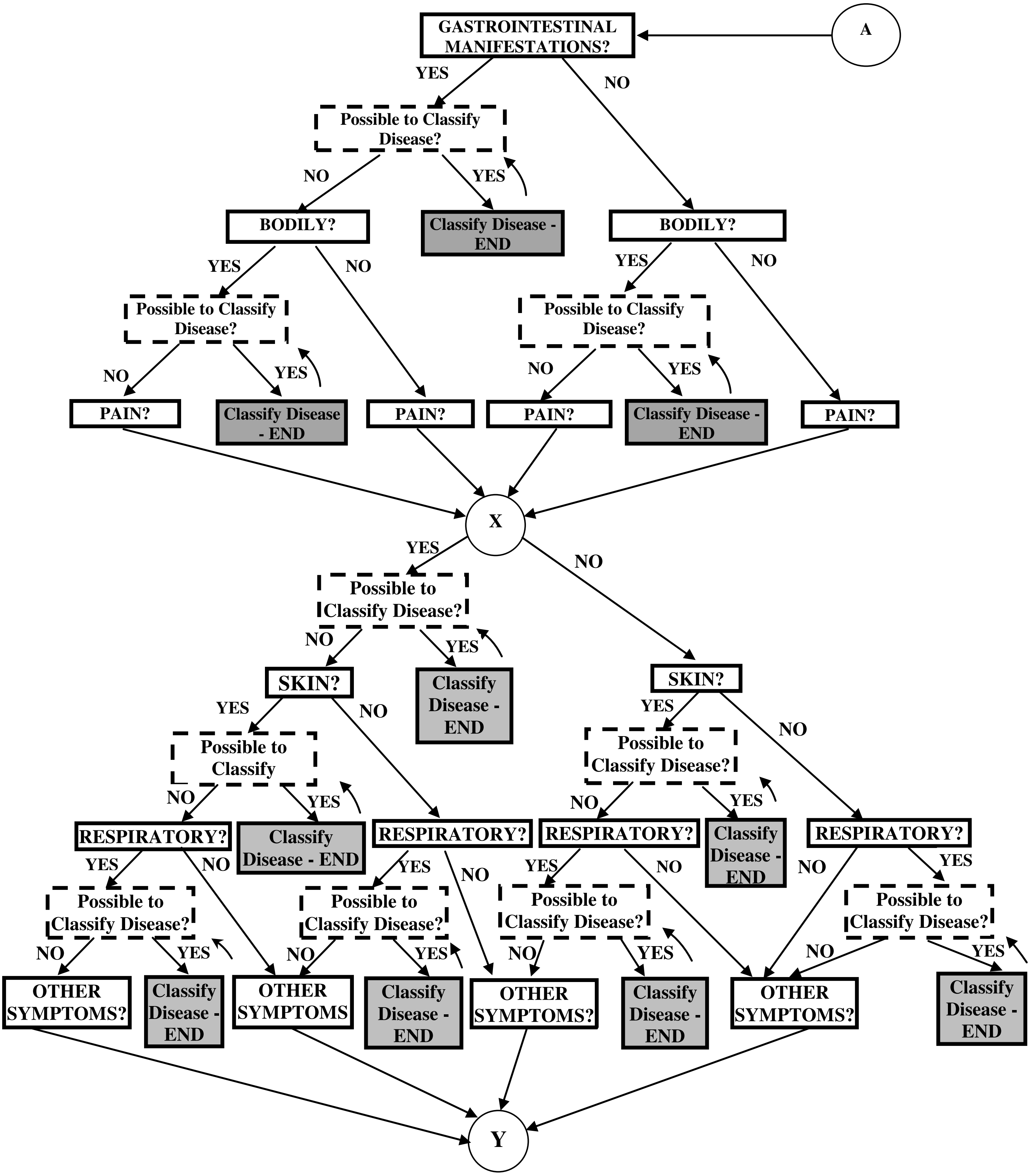


Figure 2: Decision Tree Construction (based on Information Gains Rankings of Disease Burden Variables)

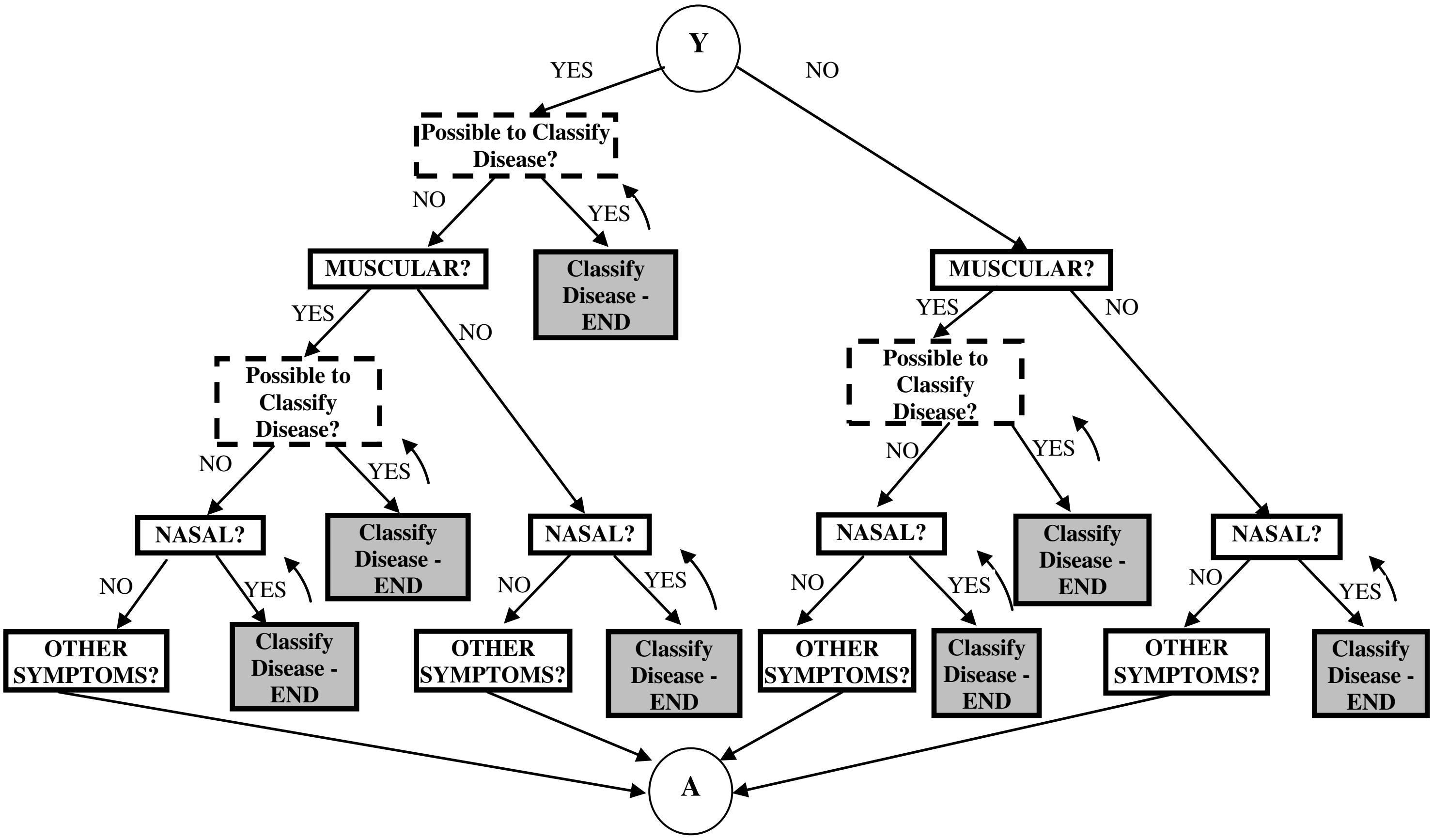


Figure 3: Decision Tree Construction (Continuation)

7. Findings And Analysis

From the results, it possible to define and construct a decision tree from any disease data set using the algorithmic model defined here. This gives the disease surveillance personnel an objective method of dynamically defining decision trees that inform and drive their surveillance operations. This can hopefully give the fight against disease outbreaks the impetus it needs since they can make better and informed ways of usefully aggregating the disease data and packaging it in a manner that helps make the most of out of it.

The research supports the collaborative efforts and co-operation between medical and information technology practitioners to create a new body of knowledge. It can only be hoped that further research on this can be extended to come up with more creative and objective methods of crunching through medical records to usefully drive decision-making and the relevant mitigating actions. Additionally, such research outcomes can also be useful in informing and driving the development of policies for both industry as well as governments in the fight against disease outbreaks.

Finally, the model developed and tested here does answer the question regarding the most critical disease burden variables to use to prioritize in predicting and fighting disease outbreaks; it variously generates the information gains and ranks for each variable to enable the construction of the desired decision tree. It also accordingly upholds the null hypothesis i.e. *the Disease Symptom Burden variables can be used to determine the information gains and consequent rankings for decision tree nodes for disease classification and prediction.*

8. Recommendations

Since the algorithm defined here is only making use of the C4.5 machine learning algorithm, it may have some limitations especially in cases in which the data assumes nominal values. In this case, it is recommneded that an alternative algorithm be defined (probably using the ID3 machine learning algorithm) to cater for cases that do not fit well with the C4.5.

The data sets used here were derived from Kenya's Nariobi County, mainly aggregated data sets for disease outbreaks for 2015 through 2018. The algorithm could be extended for other periods and regions to determine periodic disease outbreak focuses driven by the dynamically defined/derived decision trees.

9. Conclusion

With the right and objective approach, a combination of medical and informatics concepts and techniques, disease outbreak data can now be better utilized. This can also drive the government's planning and policy development 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.

10. List of References

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.

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.

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

Disease Surveillance and Response Unit (DSRU). (2014).

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

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.

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

Mayo Clinic Online Database of Diseases and their symptoms (https://www.mayoclinic.org/diseases-conditions/index)

Neill, D. B. (2012). New directions in artificial intelligence for public health surveillance. IEEE Intelligent Systems, 27(1), 56-59.

Patel, N., & Singh, D. (2015). An Algorithm to Construct Decision Tree for Machine Learning based on Similarity Factor. International Journal of Computer Applications, 111(10).

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.

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.

Russel J. R., Norvig P., (2009) Artificial Intelligence: A Modern Approach (AIMA).

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.

APPENDIX

Table 4: Kenya's Notifiable Disease List (DSRU, 2014)

Disease Code	Disease Name
AEFI	Adverse Effects Following Immunization
ATX	Anthrax
CL	Cholera
DF	Dengue Fever
DYS	Dysentery
GW	Guinea Worm
MLS	Measles
NT	Neonatal Tetanus
PLG	Plague
RVF	Rift Valley Fever
SARI	Severe Acute Respiratory Illness
VHF	Viral Hemorrhagic Fever
YF	Yellow Fever
OTH	Others