Tuberculosis Diagnosis Using Naïve Bayes Classifier

Morgan Obi, Kelechi K. Nwauzi, Sylvester Akhetuamen

Abstract — Tuberculosis is a highly contagious ailment that can be easily transmitted from person to person if not quickly and accurately diagnosed and treated. In this work, we proposed a model to diagnose this disease and take appropriate curative measure as soon as it is diagnosed. This is done using Naïve Bayes Classifier which is a supervised machine learning technique. We decided to use the Naïve Bayesian Classifier which is based on Bayes' theorem which assumes independent assumptions between predictors for this work because it assumes that the effect of the value of a predictor (x in this case a symptom) on a given class (c) is independent of the values of other predictors.

Index Terms — Bayesian, Classification, Frequency, Laplace Smoothing, Posterior, Prior, Probability, Swine Flu, Fatigue

1 INTRODUCTION

Tuberculosis (TB) is a potentially severe contagious ailment that mostly affects the lungs. The bacteria that cause tuberculosis known as Mycobacterium tuberculosis is an airborne pathogen that can easily be transmitted from person to person through tiny droplets released into the air via sneezing, coughing, singing, talking etc.

It is a chronic disease that is progressive as well as contagious. In addition to lungs, the bacteria can also infect the bones, kidneys and brain. This disease is rated as the 15th of the top 50 causes of death in Nigeria with an estimation of 29.12 death rates per 100,000 populations [1].

In this research, we proffer a way for quick and accurate diagnosis of this disease using Naïve Bayes Classifier.

2 LITERATURE REVIEW

[2] used Naïve bayes and Support Vector Machine (SVM) for liver disease prediction. Comparisons of these algorithms were done and it is based on the performance factors classification accuracy and execution time. From their experimental results, the work concludes, that SVM classifier is considered as a best algorithm because of its highest classification accuracy. On the other hand, while comparing the execution time, the Naïve Bayes classifier needs minimum execution time.

[3] used Naïve Bayes and FT Tree algorithms for disease prediction of three major liver diseases (Liver cancer, Cirrhosis and Hepatitis) with the help of distinct symptoms. They compare these two algorithms based on their classification accuracy measure. From the experimental results they concluded that Naïve bayes has the better algorithm which predicted diseases with maximum classification accuracy than the other algorithm.

[4] proposed a method to identify swine flu by studying 110 symptoms in other to decrease the cost incurred in the test of the disease. The authors developed a prototype intelligent swine flu prediction software using Naïve Bayes classifier technique for classifying the patients with swine flu. Based on the possibility of the diseases and guaranteed the accuracy of almost 63.3 percent was gotten.

3 METHODOLOGY

Questionnaires were administered to medical professionals and the result gotten is presented in table 3 below. The dataset used for training the system has 11 attributes with a summarized total of 17 instances. The attributes are cough, chest pain, bloody sputum, weight loss, drenching night sweats, fever, headache, loss of appetite, fatigue, age and HIV status.

Table 1: Attributes and abbreviation

Attributes	Abbreviation
Cough > 3 weeks	CO
Chest pain	CP
Bloody sputum	BS
Weight loss	WL
Drenching night sweat	NS
Fever	FV
Headache	HA
Loss of appetite	LA
Fatigue	FA
Age	AG
HIV status	HIV

Table 2: Severity and abbreviation (number representation)

Severity	**Abbreviation**
Mild	MI (1)
Moderate	MO (2)
Severe	SV (3)
Very severe	VS (4)

- *Morgan O. is a part-time lecturer in the Department of Computer Science, Eastern Polytechnic, Port Harcourt Rivers State, Nigeria and also a freelance software developer. E-mail: purity2m@gmail.com*
- *Kelechi K.N. is a lecturer in the Department of Electrical Electronics, Captain Elechi Amadi Polytechnic, Rivers State, Nigeria. E-mail: kkirian@yahoo.com*
- *Sylvester A. is a lecturer in the Department of Computer Science, Auchi Polytechnic, Edo State, Nigeria. E-mail: osaremhe@yahoo.com*

Table 3: Dataset of tuberculosis in Indexed format

SN	IF											THEN
	CO	CP	BS	WL	NS	FV	HA	LA	FA	AG	HIV	RESULT
1	3	0	3	3	1	3	0	3	0	0	0	VS
2	2	0	1	1	1	1	0	1	0	0	0	MI
3	0	0	0	0	0	0	3	3	0	0	2	MI
4	1	1	3	1	3	3	1	0	0	0	1	SV
5	2	0	2	2	2	2	0	2	0	0	0	MO
6	4	0	3	3	3	3	0	0	0	0	0	VS
7	4	0	1	3	2	4	0	2	0	0	0	VS
8	1	0	1	1	1	1	0	1	0	0	0	MI
9	3	0	4	1	3	3	0	3	0	0	0	SV
10	2	0	3	2	4	2	0	1	0	0	0	SV
11	4	0	2	2	2	2	0	1	2	0	0	VS
12	3	0	1	3	2	3	0	2	2	0	0	SV
13	2	0	2	1	1	2	0	2	1	0	0	MO
14	1	0	2	1	1	1	0	2	2	0	0	MI
15	1	0	2	1	2	2	0	2	1	0	0	MO
16	3	0	3	3	2	3	0	2	2	0	0	SV
17	2	0	3	3	0	2	2	0	0	0	2	MO

2.1 Naïve Bayes Classifier

The Naïve Bayesian Classifier is based on Bayes' theorem with independence assumptions between predictors. It assumes that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence.

Naïve Bayes Classifier is often used to work out posterior probabilities given observations. For example, a patient may be observed to have certain symptoms and using Bayes' theorem, the probability that a proposed diagnosis is correct, given the observation can be calculated [5].

The formula for calculating conditional probability is given below:

$$P(H/E) = P(H)\prod_{i=1}^{n} P(E_i/H)$$

Where

$P(H)$ is the probability of hypothesis H being true. This is known as the prior probability. Also known as the prior probability of class.

$P(E)$ is the probability of the evidence (regardless of the hypothesis). Also known as the prior probability of the predictor.

$P(E|H)$ is the probability of the evidence given that hypothesis is true. Also known as the likelihood which is the probability of predictor given class.

$P(H|E)$ is the probability of the hypothesis given that the evidence is true. It is also known as the posterior probability of class (target) given predictor (attribute)

Since the predictors are independent, probability of all the attributes are multiplied to get the posterior probability $P(H|E)$.

$$P(H|E) = P(e_1|h) * P(e_2 \mid h) * \ldots * P(e_n \mid h) * P(h)$$

To calculate the posterior probabilities, we first construct the frequency table for each attribute against the target using the existing tuberculosis data shown in table 3, we construct the frequency and likelihood tables for our Naïve Bayes classifier using the pseudocode shown below:

2.2 Pseudocode of Naïve Bayes Classifier

Step 1: Scan the dataset (storage servers)

Step 2: Calculate the frequency and likelihood probability of each attribute value.

Step 3: Using Naive Bayesian equation, calculate the posterior probability for each class.

Step 4: Multiply all the probabilities with respect to each attributes

Step 5: Compare the values and classify the attribute values to one of the predefined set of class variable with maximum value.

Table 4: Probability of class

P(VS)	4/17
P(SV)	5/17
P(MO)	4/17
P(MI)	4/17

Table 5: Frequency and likelihood table for Cough

COUGH	DIAGNOSIS								PRIOR PROBABILITY OF PREDIC
	VS		SV		MO		MI		
	FRQ	LKH	FRQ	LKH	FRQ	LKH	FRQ	LKH	
0	0	0/4	0	0/5	0	0/4	1	1/4	1/17
1	0	0/4	1	1/5	1	1/4	2	2/4	4/17
2	0	0/4	1	1/5	3	3/4	1	1/4	5/17
3	1	1/4	3	3/5	0	0/4	0	0/4	4/17
4	3	3/4	0	0/5	0	0/4	0	0/4	3/17

Table 5 shows that from the training data, the frequency count of cough not specified when diagnosis was very severe is 0 with a likelihood of 0/4. It also shows that the frequency count for cough specified as being mild when diagnosis was severe is 0 with a likelihood of 0/5 also the frequency count of cough specified as moderate when diagnosis was moderate is 0 with a likelihood of 0/4 while the frequency count for cough specified as mild when diagnosis is mild is 1 with a likelihood of 1/4. Adding this up gives the prior probability of cough not being specified as 1/17. Tables 6 to 8 shows frequency and likelihood values for some other symptoms. Other tables not shown can be infer from the dataset in table 3.

Table 6: Frequency and likelihood table for Chest Pain

CHEST PAIN	DIAGNOSIS								PRIOR PROBABILITY OF PREDIC
	VS		SV		MO		MI		
	FRQ	LKH	FRQ	LKH	FRQ	LKH	FRQ	LKH	
0	4	4/4	4	4/5	4	4/4	4	4/4	16/17
1	0	0/4	1	1/5	0	0/4	0	0/4	1/17
2	0	0/4	0	0/5	0	0/4	0	0/4	0/17
3	0	0/4	0	0/5	0	0/4	0	0/4	0/17
4	0	0/4	0	0/5	0	0/4	0	0/4	0/17

Table 7: Frequency and likelihood table for Bloody Sputum

BLOODY SPUTUM	DIAGNOSIS								PRIOR PROBABILITY OF PREDIC
	VS		SV		MO		MI		
	FRQ	LKH	FRQ	LKH	FRQ	LKH	FRQ	LKH	
0	0	0/4	0	0/5	0	0/4	1	1/4	1/17
1	1	1/4	1	1/5	0	0/4	2	2/4	4/17
2	1	1/4	0	0/5	3	3/4	1	1/4	5/17
3	2	2/4	3	3/5	1	1/4	0	0/4	6/17
4	0	0/4	1	1/5	0	0/4	0	0/4	1/17

Table 8: Frequency and likelihood table for Weight Loss

WEIGHT LOSS	DIAGNOSIS								PRIOR PROBABILITY OF PREDIC
	VS		SV		MO		MI		
	FRQ	LKH	FRQ	LKH	FRQ	LKH	FRQ	LKH	
0	0	0/4	0	0/5	0	0/4	1	1/4	1/17
1	0	0/4	2	2/5	2	2/4	3	3/4	7/17
2	1	1/4	1	1/5	1	1/4	0	0/4	3/17
3	3	3/4	2	2/5	1	1/4	0	0/4	6/17
4	0	0/4	0	0/5	0	0/4	0	0/4	0/17

2.3 Zero-frequency problem

The zero frequency problems occur when a particular attribute does not appear in the observation. As can be seen from our frequency and likelihood tables above, most of the attributes have zero count meaning that they were not seen in the data collected. When this happens, the result becomes zero since the result is computed by multiplying attribute occurrences. According to [6], one strategy to overcome the zero-frequency

problem could be to start all frequency counts at k = 1, rather than zero, then simply increment symptoms frequencies as symptoms are observed a method known as Laplace smoothing.

$$P(D = a) = \frac{a + k}{T + n * k}$$

Where k is the Laplace Smoothing value (1), a is the attribute value, T is the total number of class attributes and n is the number of class.

2.4 Laplace Smoothing Applied to frequency and Likelihood Table

The following tables show the frequency and likelihood values after Laplace smoothing was applied on our training dataset.

Table 9: Probability of Class after Laplace Smoothing

P(VS)	5/21
P(SV)	6/21
P(MO)	5/21
P(MI)	5/21

Table 10: Frequency and likelihood table for Cough after Smoothing

COUGH	DIAGNOSIS								PRIOR PROBABILITY OF PREDIC
	VS		SV		MO		MI		
	FRQ	LKH	FRQ	LKH	FRQ	LKH	FRQ	LKH	
0	1	1/5	1	1/6	1	1/5	2	2/5	5/21
1	1	1/5	2	2/6	2	2/5	3	3/5	8/21
2	1	1/5	2	2/6	4	4/5	2	2/5	9/21
3	2	2/5	4	4/6	1	1/5	1	1/5	8/21
4	4	4/5	1	1/6	1	1/5	1	1/5	7/21

Table 11: Frequency and likelihood table for Chest Pain after Smoothing

CHEST PAIN	DIAGNOSIS								PRIOR PROBABILITY OF PREDIC
	VS		SV		MO		MI		
	FRQ	LKH	FRQ	LKH	FRQ	LKH	FRQ	LKH	
0	5	5/5	5	5/6	5	5/5	5	5/5	20/21
1	1	1/5	2	2/6	1	1/5	1	1/5	5/21
2	1	1/5	1	1/6	1	1/5	1	1/5	4/21
3	1	1/5	1	1/6	1	1/5	1	1/5	4/21
4	1	1/5	1	1/6	1	1/5	1	1/5	4/21

Table 12: Frequency and likelihood table for Bloody Sputum after Smoothing

BLOODY SPUTUM	DIAGNOSIS								PRIOR PROBABILITY OF PREDIC
	VS		SV		MO		MI		
	FRQ	LKH	FRQ	LKH	FRQ	LKH	FRQ	LKH	
0	1	1/5	1	1/6	1	1/5	2	2/5	5/21
1	2	2/5	2	2/6	1	1/5	3	3/5	8/21
2	2	2/5	1	1/6	4	4/5	2	2/5	9/21
3	3	3/5	4	4/6	2	2/5	1	1/5	10/21
4	1	1/5	2	2/6	1	1/5	1	1/5	5/21

Table 13: Frequency and likelihood table for Weight Loss after Smoothing

WEIGHT LOSS	DIAGNOSIS								PRIOR PROBABILITY OF PREDIC
	VS		SV		MO		MI		
	FRQ	LKH	FRQ	LKH	FRQ	LKH	FRQ	LKH	
0	1	1/5	1	1/6	1	1/5	2	2/5	5/21
1	1	1/5	3	3/6	3	3/5	4	4/5	11/21
2	2	2/5	2	2/6	2	2/5	1	1/5	7/21
3	4	4/5	3	3/6	2	2/5	1	1/5	10/21

4	1	1/5	1	1/6	1	1/5	1	1/5	4/21

4 RESULTS AND DISCUSSION

The Naïve Bayes model built was evaluated using the patient symptoms shown in table 14 which represents a patient having very severe cough, with no value indicated for chest pain, a moderate bloody sputum, mild weight loss, moderate night sweat, moderate fever, nothing indicated for headache, severe loss of appetite, and nothing indicated for fatigue, patients age and HIV status:

Table 14: A patient presenting symptoms for diagnosis

SN	CO	CP	BS	WL	NS	FV	HA	LA	FA	AG	HIV	RESULT
1	4	0	2	1	2	2	0	3	0	0	0	?

Using our trained model, the diagnosis for the patient can be done as follows:

Using the Naïve Byaes formula to calculate the conditional probability given below:

$$P(H|E) = \frac{P(E|H) * P(H)}{P(E)}$$

Since the predictors are independent, probability of all the attributes are multiplied to get the posterior probability $P(H|E)$.

$$P(Diagnosis|Symptoms) = P(e_1|h) * P(e_2 \mid h) * \dots * P(e_n \mid h) * P(h)$$

$$P(VS) = P(CO = 4/VS) * P(CP = 0/VS) * P(BS = 0/VS) * P(WL = 1/VS) * P(NS = 2/VS) * P(FV = 2/VS) * P(HA = 0/VS) * P(LA = 3/VS) * P(FA = 0/VS) * P(AG = 0/VS) * P(HIV = 0/VS) * P(VS)$$

$$P(SV) = P(CO = 4/SV) * P(CP = 0/SV) * P(BS = 0/SV) * P(WL = 1/SV) * P(NS = 2/SV) * P(FV = 2/SV) * P(HA = 0/SV) * P(LA = 3/SV) * P(FA = 0/SV) * P(AG = 0/SV) * P(HIV = 0/SV) * P(SV)$$

$$P(MO) = P(CO = 4/MO) * P(CP = 0/MO) * P(BS = 0/MO) * P(WL = 1/MO) * P(NS = 2/MO) * P(FV = 2/MO) * P(HA = 0/MO) * P(LA = 3/MO) * P(FA = 0/MO) * P(AG = 0/MO) * P(HIV = 0/MO) * P(MO)$$

$$P(MI) = P(CO = 4/MI) * P(CP = 0/MI) * P(BS = 0/MI) * P(WL = 1/MI) * P(NS = 2/MI) * P(FV = 2/MI) * P(HA = 0/MI) * P(LA = 3/MI) * P(FA = 0/MI) * P(AG = 0/MI) * P(HIV = 0/MI) * P(MI)$$

According to the pseudocode, our diagnosis will be the class variable with maximum value.

$$P(VS) = (4/5) * (5/5) * (1/5) * (1/5) * (3/5) * (2/5) * (5/5) * (2/5) * (4/5) * (5/5) * (5/21)$$

$$P(VS) = 0.8 * 1 * 0.2 * 0.2 * 0.6 * 0.4 * 1 * 0.4 * 0.8 * 1 * 0.24$$

$$P(VS) = 0.00059$$

$$P(SV) = (1/6) * (5/6) * (1/6) * (3/6) * (3/6) * (2/6) * (5/6) * (2/6) * (4/6) * (5/6) * (6/21)$$

$$P(SV) = 0.167 * 0.833 * 0.167 * 0.5 * 0.5 * 0.333 * 0.833 * 0.333 * 0.667 * 0.833 * 0.286$$

$$P(SV) = 0.000085$$

$$P(MO) = (1/5) * (5/5) * (1/5) * (1/5) * (3/5) * (5/5) * (4/5) * (1/5) * (3/5) * (4/5) * (5/21)$$

$$P(MO) = 0.2 * 1 * 0.2 * 0.2 * 0.6 * 1 * 0.8 * 0.2 * 0.6 * 0.8 * 0.24$$

$$P(MO) = 0.000088$$

$$P(MI) = (2/5) * (5/5) * (2/5) * (2/5) * (2/5) * (1/5) * (4/5) * (2/5) * (4/5) * (4/5) * (5/21)$$

$$P(MI) = 0.4 * 1 * 0.4 * 0.4 * 0.4 * 0.2 * 0.8 * 0.4 * 0.8 * 0.8 * 0.24$$

$$P(MI) = 0.00025$$

Haven gotten the likelihood of the severities, the probability of the severity can be obtained by normalizing the result. This is done using the formula below:

$$prob(severity) = \frac{likelihood\ of\ severity}{\sum severities\ likelihood}$$

Probability of Very Severe

$$prob(vs) = \frac{0.00059}{(0.000085 + 0.000088 + 0.00025 + 0.00059)}$$

$$prob(vs) = \frac{0.00059}{0.001013} = 0.5824$$

Probability of Severe

$$prob(sv) = \frac{0.000085}{(0.000085 + 0.000088 + 0.00025 + 0.00059)}$$

$$prob(sv) = \frac{0.000085}{0.001013} = 0.00839$$

Probability of Moderate

$$prob(mo) = \frac{0.000088}{(0.000085 + 0.000088 + 0.00025 + 0.00059)}$$

$$prob(vs) = \frac{0.000088}{0.001013} = 0.0869$$

Probability of Mild

$$prob(mi) = \frac{0.00025}{(0.000085 + 0.000088 + 0.00025 + 0.00059)}$$

$$prob(vs) = \frac{0.00025}{0.001013} = 0.2468$$

Based on our training data and computation, the presented data shows that:

$$P(VS) > P(MI) > P(MO) > P(SV)$$

Hence, we can conclude that the patient with symptoms given as shown in table 14 have a TUBERCULOSIS case that is VERY SEVERE and requires urgent medical attention and needs to be quarantined to curtail spreading of the disease.

5 CONCLUSION

In this work, we designed a model for tuberculosis diagnosis. Our dataset was gathered from trained and practicing medical practitioners and the dataset was trained using Naïve Bayes Classifier. The model designed was used to diagnose some test dataset and the output matches with that of trained professionals.

REFERENCES

[1] WHO. (2016). Chest Radiography in Tuberculosis Detection. Switzerland: Knut Lönnroth.

[2] Vijayarani1 S, Dhayanand S. Liver Disease Prediction using SVM and Naïve Bayes Algorithms International Journal of Science, Engineering and Technology Research (IJSETR) Volume 4, Issue 4, April 2015

[3] Dhamodharan. S, Liver Disease Prediction Using Bayesian Classification, Special Issue, 4th National Conference on Advanced Computing, Applications & Technologies, May 2014, page no 1-3.

[4] Thakkar. Hasan and Desai, "Health care decision support system for swine flu prediction using Naïve bayes classifier", International Conference on Advances in Recent Technologies in Communication and Computing., IEEE, 2010.

[5] Pattekari, S. A., & Parveen, A. (2012). Prediction System for Heart Disease using Naive Bayes. International Journal of Advanced Computer and Mathematical Sciences, 290 - 294.

[6] Steven, P. (2014, September 23). The zero frequency problem (Part 1). Retrieved October 1, 2017, from Harder, Better, Faster, Stronger: https://hbfs.wordpress.com