Response and Surveillance System for Diarrhoea Based on a Patient Symptoms Using Machine Learning: A Study on Eswatini

SIBONGAKONKE KWANELE ZUNGU1, QI-XIAN HUANG1, MIN-YI CHIU1, AND HUNG-MIN SUN2
1Institute of Information Systems and Applications, National Tsing Hua University, Hsinchu 30013, Taiwan
2Department of Computer Science, National Tsing Hua University, Hsinchu 30013, Taiwan
Corresponding author: Sibongakonke Kwanele Zungu (s108065432@m108.nthu.edu.tw)

This work was supported in part by the Ministry of Science and Technology, Taiwan, under Project MOST 107-2221-E-007-015-MY3 and Project MOST 110-2221-E-007-040-MY3; and in part by Senao Network Corporation.

ABSTRACT Utilizing supervised machine learning algorithms to develop a surveillance and response system based on symptoms of diarrhoea, contingent on the Support Vector Machine (SVM) to predict the probable disease using labelled data. Diarrhoea is amongst the top ten diseases which kill. A prototype system is developed based on the SVM algorithm. The prototype system takes six patient symptoms that which is input, from the user and the output result becomes the prognosis which may likely occur based solely on the given symptoms. Two other supervised learning models have been utilized in the prediction process, Random Forest Model (RFC) and Naïve Bayes Model (NB). Furthermore, a visualization on google maps (my maps) on the area in which a diarrhoea outbreak would likely occur. The constituency and the region of the patient will be used to place a pin on my maps, giving a visualization on the map, with a mapping structure this allows for a vivid demonstration of how diarrhoea is spreading in Eswatini. SVM received an average of 100% accuracy. The other two supervised learning models, random forest model and naïve Bayes model received 97.62% average accuracy on the same dataset. It shows that the SVM does well in data classification and with a small dataset.

INDEX TERMS Diarrhoea, prognosis, supervised machine learning.

I. INTRODUCTION

This research explores ways in which acute diarrhoea can be detected at early stages within communities in Eswatini to widely reduce the chances of an outbreak, mostly in children under the age of 5. When it comes to experiencing acute diarrhoea [7] it should be short-lived. When acute diarrhoea spans over weeks, there is a major concern.

The World Health organization defines acute diarrhoea “as the passage of three or more loose or liquid stools per day”, [7]. There has been nearly 1.7 billion annually recorded global cases of childhood diarrhoeal diseases [7]. A Global Burden of Disease Study which was conducted in 2017 showed that, there have been 22167 total deaths due to diarrhoea for the past 27 year, from 1990 to 2017. The death toll of children under 5 has been 12454 in these years.

This alludes that 56.2% of the whole population in which had diarrhoea were children under the age of 5. Figure 1 shows the trend of deaths of children under the age of 5. Looking at figure 1, we see that over the years there have been high cases of diarrhoea, however the cases are slowly decreasing. With that, there is still a need to closely monitor diarrhoea to ensure that there is no potential outbreak. There are three clinical types of diarrhoea defined by the World health organization, which are:

• Acute watery diarrhoea: This type of diarrhoea lasts for several hours or days [7].
• Acute bloody diarrhoea: It is also called dysentery [7].
• Acute Persistent diarrhoea: This type last for 14 days or more [7].

In addition, diarrhoeal diseases are amongst the top ten fatal diseases. Diarrhoeal diseases come second in children under the age of 5 mortality and it claims around 525 000 lives of children under the age of 5 annually [7].
With the alarming rates of deaths of children under the ages 5 globally, there has been measures by which have been taken to reduce the detrimental effects of diarrhoea. Vaccination has been one of the preventative measures which have been put in place.

In order to effectively develop an effective algorithm which will be implemented in the prototype system, there is a need to review different papers which will firstly give an overview of diarrhoea at an international level and further narrow it down to a national level in Eswatini. Hence papers used in the study have been divided into two groups: the first group of papers are papers which discuss issue of diarrhoea, such as how it spreads from one person to the other. Secondly, diarrhoea being a communicable disease and ways in which diarrhoea can be controlled or eventually eradicated. Lastly, the papers further highlight places which may be prone to a diarrhoea outbreak due to certain circumstances such as the type of settlements in which the people stay, the availability of sanitary water and clean lavatories. These three factors are covariates of diarrhoea.

The second group of papers highlight the different algorithms which different scholars have utilized to predict a disease or give a prognosis, based on given symptoms. There were several different algorithms which were used in the different papers by different scholars. Algorithms for this research were chosen based on their ability to:

1) Effectively train on a small dataset with many features.
2) Quick prediction of prognosis from the provided symptoms, a model with a high training time is needed.
3) Good with classification. My dataset is more of a text dataset, hence I need algorithms which do better with a text dataset.
4) High accuracy as it is vital, the correct prognosis is predicted based on the symptoms given to the system.

The supervised learning algorithms which are able to meet the above listed criteria are NB, SVM and RFC. A comparison of the three different models will be conducted based on their abilities to correctly predict diarrhoea.

Diarrhoea is still a problem in Eswatini, hence constant surveillance of diarrhoea is vital. This will help detect diarrhoea early and minimize its effect. The research also aims at creating a database of symptoms and diseases different individuals experience from different ages and locations in Eswatini. This database can further help develop a new system of finding a correlation between a person’s location and symptoms being experienced.

A. PROBLEM STATEMENT
Despite the roll out of a diarrhoea vaccine in Eswatini, there are still cases of children under the age of 5 suffering from diarrhoea in the country, hence the importance of strong monitoring of cases that arise in order to make sure there is no future outbreak [16].

B. AIM
With current state of the world, Covid-19 has placed pressure on the health facilities and departments, a tool is needed to further assists fight the equally life-threatening diseases like diarrhoea in the country. With a system which is based on an algorithm which takes symptoms of patients and give
a prognosis will further help the health sector. This will help with keeping track of the different symptoms people have in a specific region, the age rage of the people with those symptoms and this will allow the health sector to act promptly.

C. ASSUMPTION
There are factors which have not been considered in this study such as, patience previous illnesses and family history diseases. The prediction of diarrhoea will be solely based on the six symptoms which should be given as input and the output will be the probable disease which might result in the prediction. In the case of this research it will be diarrhoea.

D. MOTIVATION
There has been vast number of children who die due to diarrhoea globally. There are 1.7 billion annually reported case of children who have had diarrhoea. Diarrhoea is a treatable and preventable disease. This means that there are measures which can be taken to prevent and monitor the disease with the aim to reduce the large number of children who are affected globally [7]. The ultimate motivation is to develop a system that will monitor and rapidly respond to a potential diarrhoea outbreak and to help take necessary precautions in alerting the different health sectors on what is happening in the different regions of the country as an attempt to reduce fatality of diarrhoea amongst children under the age of 5.

E. RESEARCH QUESTIONS
1) What is the most efficient algorithm for diarrhoea surveillance?
2) What are the best ways to find detect the spread of a disease?
3) What are the major causes of diarrhoea?
4) How does diarrhoea spread?
5) Which region is mostly affected and why?

F. OBJECTIVES
- To efficiently predict a potential diarrhoea spread within a community
- To evaluate and compare effective models for disease prediction based on patient symptoms.
- To provide a method or tool of action and decision making
- To provide rural or remote areas with an effective prediction tool
- To find places which are prone to diarrhoea based on unsanitary water as one of the covariate.

G. CONTRIBUTION
Machine learning algorithms have been widely used in the medical field as a tool for assistance in predicting a disease based on different variables such as patient family health history and patient disease symptoms. This research contributes to:

1) The knowledge that a correctly predicted prognosis can help develop a system which can monitor diseases to help prevent a potential outbreak.
2) To further collect and store data to be used for future analysis.

The expected result of the research is a highly effective algorithm which correctly predicts diarrhoea based on major six symptoms of diarrhoea and further vividly plotting the cases that arise on the map of Eswatini using the location provided by the patient when providing their details.

This paper is divided into various sections and subsections. In section I, we have the introduction which encapsulates the problem statement, aim, assumption, motivation of the study, some research questions and lastly the objectives. In section II a review of other scholar articles related to diarrhoea are discussed. In section III the proposed methodology is discussed, under which the different tools used for the proposed prototype system will be highlighted. In section IV, the results of the study are presented and a comparison of the three models is done. In section V which gives a brief discussion of the paper, concludes the paper with an entire overview of the experiments conducted and state some of the limitations of the study. Lastly in section VI we have the paper acknowledgements.

II. RELATED WORK
According to the world health organisation (WHO) in an article written in 2017 stating that diarrhoea is the amongst the top ten causes of death in children under the age of 5, it is rated second on the list of fatal diseases. There have been vast number of children who die due to diarrhoea globally. Reported cases about children who have had diarrhoeal diseases are at 1.7 billion annually. We note that acute diarrhoea is a disease that both preventable and treatable. This further suggests that there are several measures which can be taken to prevent and monitor diarrhoea with the aim to reduce the large number of childhood deaths, especially those who are under the age of 5.

The world health organization further links the causes of acute diarrhoea to poor sanitation and hygiene, the unavailability or poor state of lavatories, the lack of clean water for drinking, cooking and cleaning. Kids who suffer from malnutrition are more vulnerable to being affected by diarrhoea. With the lack of the mentioned essential variables to maintain a good health life, diarrhoea bacteria are manifested in these surroundings and major symptoms are loose stool and a more frequent visit to the lavatory in a day than the usual. The most important thing is not to provide breeding ground for these parasitic organisms as it can spread from one person to another through unsanitary practices [7]. It has been noted that in low-income countries there is usually more cases of diarrhoea as there is a correlation with water sanitation.

The world is changing at a rapid speed and it has become a challenge to keep up with the changing times. Creating a good health and well-being society has been stated as one of Eswatini’s sustainable development goals (SDG’s) which was
stated by United Nations Eswatini 2030 epidemics of diseases mentioning diseases such as cholera, tuberculosis and other communicable and non-communicable diseases should be eradicated or under immense control. Hence the importance of surveillance of diseases to increase a preventative measures and alertness of diseases which may cause a potential diarrhoea spread in the country.

Unsanitary water and the unavailability of lavatories in some parts of the country have caused major obstacles in Eswatini, where there is a high rate of people dying due to diarrhoea. The unsanitary water is not only consumed through drinking, however, it is used for multiple purposes such as: watering gardens, cooking and the like. Water Aid states that over 200 children who are under the ages of 5 die from diarrhoea in Eswatini. It is evident that children under the age of 5 are the most affected by diarrhoea. There have been efforts made in the past and are still in place to improve water sanitary and the availability of lavatories in the country.

According to an article written by Maphalala G, the author mentioned that the peak months were between June and August which is usually the winter season. These are the dry months throughout Eswatini, this means that there is less water in the country and some people in the rural areas get water from the rivers. There is a need to go through the process of cleaning the water before it can be used for both cooking and drinking.

The most common type of diarrhoea observed in Eswatini is acute watery diarrhoea. A 2014 outbreak prompted the current vigilance and the vaccination which was introduced in May 2015 as stated by the Epidemiology and Disease Control Unit (EDCU) under the Eswatini Diarrhoea Contingency plan 2018. Ever since the 2014 outbreak, there has been a system in place to monitor diarrhoea to make sure that there is no outbreak which will go undetected.

There is an association between improved water supply or improved water sanitation and garbage disposal with diarrhoea. In an article written by Subtasks, a study conducted in Ethiopia, it was mentioned that an improvement of access to clean water and sanitation services and the cause of diarrhoea was the main objective of the paper. It was presented in the paper that a cross-sectional study was conducted using data from an Ethiopian 2016 Demographic health survey [4].

Mothers or caregivers of children under the age of 5 were interviewed and a logistic regression analysis was then utilized to examine the relationship between the different variables which are independent and dependent. It was discovered that having more than 2 children under the age of 5 under one household which also has 5 members per household and the way garbage was disposed were strongly associated [4]. Hence reducing the crowding within houses and practicing safe garbage disposal would reduce the cases of diarrhoea [4].

III. PROPOSED METHODOLOGY

In this proposed method of surveillance and response system focusing on acute diarrhoea symptoms among other diseases, the aim is to use three supervised machine learning models and to use google maps (my maps) which will be used for location purposes on the map of Eswatini to map out the potential outbreak of diarrhoea, see figure 2. With this, a prototype system will be designed and will take:

1) Patient symptoms
2) Patient information where age and location are an important part of the system.

The above two variables are most important as they play a crucial role in detecting a potential outbreak in a region which has been flagged based on the shortage of water in the region constituency, availability of lavatories and type of settlement. Individuals which are mostly affected in the country are children under the age of 5. For the second variable, diarrhoea is more prominent in dry regions where they have less water. The season needs to be taken into consideration. Diarrhoea is more common in the dryer months of the year in Eswatini between June and August, these months span over the winter and beginning of spring seasons. [9].

The three models which will be modelled on the same training and testing dataset which constitute of seen and labelled data. The algorithm which has a high accuracy and predicts diarrhoea correctly will be used for the prototype system. The models will be designed such that the system can be modified for other diseases to prediction in the future which will be based on the symptoms given to the model. A comparison between the three algorithms will be performed.

A. ALGORITHM AND SYSTEM DESIGN

In this subsection we look at the overall flow of the algorithm. From the initial stags which is preprocessing all the way to the prediction stage and the accuracy stage. A design on how the algorithm will do the diagnosis prediction based on the symptoms of patients in the given dataset. The aim of the system is to find evidence that there might be a potential outbreak of diarrhoea and to find the evidence we need mining tools and from the evidence we find knowledge and thus informed decisions can be made. An action can be taken promptly to curve the potential spread of diarrhoea.

1) DATASET

The dataset which has been utilized in this research is one which is labelled, sourced from kaggle [14]. It contains symptoms which correspond to a prognosis (disease). There are two dataset files which are both labelled data which are namely training set and testing set. Depending on the type of algorithm being used the data needs to encoded before being used to train the model. In this study three supervised machine learning algorithms have been utilized, these are random forest, naive bayes classifier and support vector machine. Only in the rfc and nb will the label encoder be used in the dataset, on both the training set and the testing set, see figure 9.

The training dataset contains 133 symptoms and 5040 prognosis, that is 133 columns and 5040 rows. The training set however has the prognosis repeated and that eventually created 5040 rows of data. A sample of the datasets
Collect data containing the different symptoms of patients and the prognosis. A dataset made from ones and zeros. One showing that the symptom is present in that disease and zero showing that the symptom is not present.

Data is in two parts: Training set and label set.

Choose different machine learning models which work well with a small dataset, good in classification and which have a short run time and they are conservative with space.

Choose different machine learning models which work well with a small dataset, good in classification and which have a short run time and they are conservative with space.

Choose one algorithm with the highest accuracy to use for the prototype system which has a user interface.

Record all the information on a database which includes a prediction and other information about the patient.

Use the stored information from the database to plot data on google maps (my maps) to visually show the trends and location with the most cases of diarrhoea. The study focuses on children under the age of 5.

FIGURE 2. A flow chart showing the proposed methodology of the study.

is seen below figure 4, 5. We note that both figure 4 which is the training set and figure 5 which is the test set have zeros and ones under each column and corresponding to each prognosis. We take note that the original dataset was developed using finding associated concepts with text analysis [6] as seen in figure 3

• one(1) means: The symptom is present in the prognosis. (Active symptom)
• zero(0) means: The symptom is not present in the prognosis. (Non-active symptom).

The testing set has 133 symptoms and 42 rows for prognosis. figure 5 is which is used to test the models. The three models which are Random forest, support vector machine and naive bayes. All the three models will be trained and tested on the same dataset.

2) ALGORITHM IMPLEMENTATION

In this section we look at how the support vector machine algorithm was implemented. The value of C parameter is 7.0. This value is set to 7.0 to avoid misclassification during testing and training. If there are large values of C this means that the optimization will have to choose a smaller hyperplane to correctly classify training points. It is important that the model is as accurate at possible as this is a crucial prediction, predicting the prognosis based on symptoms. For the prototype system which is based on the SVM algorithm will only six symptoms as input.

The support vector machine algorithm works well with both linear and non-linear data. In the case of the research we are working with linear data, hence the kernel will be set to a linear kernel [1]. The dataset which is available is limited hence the algorithm which can be used has to be one which can be able to handle a small dataset and effective with high dimensional data. High accuracy is important it is also important to have less training time and using less resources is vital as resources can be scarce hence it is important that resources are well managed.
The prototype system is designed to take six symptoms as input, hence the top six most prominent symptoms of acute diarrhoea are used, these are loose stool, dizziness, fatigue, stomach pain, abdominal pain and dehydration. Looking at the table 1 below, we see how many times the six important features occur in the database. This means that some other diseases have these symptoms. When processing the training dataset, the was no feature reduction performed on it. The reason is such that when we reduce the features we will lose features like loose stool. We see loose stool only occurs once in the dataset see table 1, hence it could be viewed as a non-important feature however it is one of the major symptoms of acute diarrhoea. The support vector machine predicts the correct prognosis/expected prognosis which is diarrhoea as it has the ability to deal with high dimensional data. Figure 6 show how many times the top six diarrhoea symptoms appear in the dataset. We note that some symptoms can appear more than once as some diseases have similar symptoms to other diseases.

TABLE 1. Table showing the top six causes of diarrhoea and the number of times they occur in the dataset.

| Symptom      | Occurrence in dataset |
|--------------|-----------------------|
| loose stool  | 1                     |
| dizziness    | 18                    |
| fatigue      | 3                     |
| stomach pain | 10                    |
| abdominal pain | 2                   |

B. ALGORITHM FLOW

In this part of the thesis we look into how the algorithm will flow, from the preprocessing stage, training and testing and prognosis prediction together with attaining model accuracies. We see figure 7 which shows the three important points of the research and how information will be derived for the points. These three points are very important in the research as deriving evidence in order to create knowledge about diarrhoea from the smallest communities in remote areas, to the entire country. Once knowledge has been created through evidence, then a decision can now be taken and action can be put in place.

TABLE 2. Table showing the contents of each of the variables used in both the training and testing.

| Variable | Variable content |
|----------|------------------|
| X        | 5040 Prognosis   |
| y        | 133 symptoms     |
| X_test   | 42 Prognosis     |
| y_test   | 133 symptoms     |

Figure 8 depicts how the model of the three different models will be structured. Preprocessing techniques using a label encoder which encode target labels with values between 0 and n_classes1. Both the training set and the testing set were divided into X and y. The label encoder was only applied to y in the training data set and y_test in the testing set. The label encoder is applied both to the random forest and the naive bayes algorithm. See table 2.

We further look at figure 9(a) which shows random forest algorithm flow and figure 9(b) which shows the support vector machine. These two supervised machine learning algorithms have been used. Comparing the two we see that before we train the model we need to label encode the training set and testing set for the algorithm to be effective. If the label encoder is applied on the support vector machine, the accuracy of the model changes from being 100% hence the label encoder was not applied.

After the label encoder is applied on the dataset which is used to train the random forest algorithm and naive bayes the model the model accuracy is collected see figure 17 image (a). After training is done, the label encoder is applied to the testing algorithm and both the algorithms are tested on the same dataset. Two results are expected, the correct prediction, which is based on the input symptoms, we expect the resulting prognosis to be diarrhoea. The second result which is expected at the end of the model is the model testing accuracy.

With figure 9 we see the support vector algorithm without the label encoder like the other two algorithms discussed above. The same dataset is also sued for the SVM. The algorithm is first trained using the training set. The training accuracy is then recorded, see figure 9(a). After the model...
has been trained and accuracy is recorded then the model is tested and two outputs are expected, which are the correct prognosis based on the symptoms provided, in this case it has to be diarrhoea and lastly the accuracy after the model is tested is also recorded.

C. SYSTEM DESIGN

1) REQUIREMENT ANALYSIS

Part of the thesis contains a prototype system which gives a prediction of diarrhoea, hence it is important that requirements of the system are discussed and defining the expectations of the prototype system is vital. In this chapter we discuss some of the forms of requirement analysis which have been used which are: Scenario based diagram and entity-relationship Diagram (E-RD). We will further discuss the functional and non-functional requirements of the prototype system.

Functional Requirements
- Enter Patient Information and Save first.
- Select Symptoms from drop down list and save. System will pass an error if less than six symptoms are selected.
- Health Motivators will be the ones using the system for their communities.
- Users need to refresh before the next prediction is done.
- Google maps (my maps account) to see case location
- Running database at all times

Non-functional Requirements
- Hardware will be a necessity.
- Internet Connection will be vital.

2) SOFTWARE TOOL USAGE

- Python (Jupyter)- It was used to develop all three machine learning algorithms
- Tkinter- Software is based python, it was used for the purposes of the user interface
- Random Forest Classifier and Naive Bayes Algorithm Flow
- Support Vector Machine Algorithm Flow
- PostgreSQL-Used to create the Data Storage for the data which will be entered through the user interface Tkinter. Structured query language will be used to query patient information per need
- Microsoft Excel- Used to store converted data from PostgreSQL in csv format to be further used in my maps
- Google Maps (My maps) - Used to plot locations of cases which have been filtered in PostgreSQL and saved in csv format.

D. SCENARIO BASED DIAGRAM

In this section we will look at how the system will interact with the user and the system administrator who will be in charge of maintenance of the system. We look at how the user in our case it is the health motivators or local small clinics who will be using the system in the different communities which will further help the national health system know more about the diarrhoea conditions in the more remote areas of the country.

Figure 10 illustrate how the user of the system will operate it. The user needs to enter the patient information thereafter they need to save the information on a database.
After the information has been saved on the database, refer to figure 12 for entity relationship diagram representing the database the user will then ask the patient about their symptoms, they need to take the top six symptoms the patient gives and pass it through the system.

The symptoms of the patient will be saved in a different table named Patient_symptom, 12 with the patients national identity document as the primary key of the symptoms table and the foreign key as well. The prediction button 22 on the system predicts and saves the prediction in a separate table named Disease_Prediction 12. It is vital that the data is saved for future reference in Eswatini. Figure 11 shows the layout of the system. Figure 22 shows that after the model is trained, the test will be applied on the model and then the six symptoms can be selected and the prediction can be done by the support vector machine algorithm under classification. Once the user is done using the system, they can simply exit the page.

E. ENTITY-RELATION DIAGRAM
An Entity relation diagram in figure 12 has been designed to show the relationship between the tables with the database name diarrhoea. In the diagram, a list of attributes and their datatype are shown which will be required from the user in order to effectively save the data. In figure 12 we see the three important tables which will hold different information and the main relationship between the tables is national_id. This is the most important attribute in the table as this is how patients will be uniquely identified and it will be used for query purposes.
F. SOFTWARE TESTING

In this section we are going to look at how the prototype system was tested. We look at the test cases in the table below. It is important to have test the software to see if it does what it has promised to do.

The first test case is seen in table 3. This is where I test if the patient information is being entered on the user interface is successfully saving in the database. In figure 13 we see the data being save in the database and national_id being the primary key for the whole table.

TABLE 3. Prototype system test case 1.

| Test case 1 | enter patient information in the text boxes |
|-------------|--------------------------------------------|
| Precondition | 1. python kernel is running(Application is running) 2. Database is available to save |
| Testing steps | 1. Input patient information 2. Select the save button |
| Expected result | Message box notifying that data has been successfully saved |

In table 4 we test if the prototype systems can effectively select the six symptoms of diarrhoea and save them in the database. Table 4 also makes use of national_id which serves as the primary key of the table. In figure 14 the result of selecting the data from the six options in the option menu on the user interface.

Test case 3 looks into the functionality of the predict button in the prototype system. we see the results in figure 22 of the predict button successfully predicting diarrhoea.

IV. RESULT DISCUSSION

In this section of the paper we discuss the three different models, RFC, SVM and NB which have been trained and tested on the same data. We look at how these models have performed individually, in terms of their accuracy both on the testing set and the training set, we also study their statistical performance, the F-1 score and precision under the classification report. Lastly we will look into the correlation heat map matrix which shows either a positive or negative correlation between two symptoms.

A. EXPERIMENT RESULTS/FINDINGS

The aim of the research is to predict diarrhoea from given dataset of patient symptoms. These are general symptoms patients experience for different diseases. In section III we discussed the arrangement of the dataset and we mentioned that the supervised machine learning algorithm and labelled dataset will be used. With the use of three different supervised machine learning algorithms, we were able to achieve 100% model accuracy with support vector machine, which further accurately predicts diarrhoea given six different symptoms, see figure 22. The two other machine learning models both had a 97.62% accuracy. These were the naive bayes and the random forest Classifier.

Furthermore, in the research making use of the prototype system which is based on python and using the user interface tkinter to collect patient information and storing the information on a database. The stored data will allow the health department to query the data using structured query language to find the important aspect of the research which is the number of diarrhoea cases which have been predicted by the support vector algorithm. Second important aspect is the number of children under the age of five who have been recorded and lastly which region and constituency have both children under the age 5 and symptoms have been recorded and predicted to be that of diarrhoea.

1) SVM

For the algorithm to be effective, we need to find the ideal hyperplane that will differentiate between the two classes. In figure 15 we note how best we can perform a classification and further make a prediction. The support vector machine
With the support vector model the accuracy remained consistent. We take note that the support vector machine achieved the highest accuracy of the two models which was a 1.0 average on the testing set. We have the testing accuracies shown on a bar-graph for the different algorithms in figure 9 (b).

B. CLASSIFICATION REPORT

In the classification report we look into the F1 score, precision and recall for the three algorithms used in the study, random forest classifier, support vector machine and naive bayes. These matrices are used to measure the quality of predictions [10] from the three classification algorithms we are studying in this paper. It is important that the three different matrices are calculated to allow better comparison of the three classifiers which have been used in this thesis.

1) PRECISION

This matrix is the accuracy of positive predictions. There are different ways to check if the predictions are wrong or right [10].

1) TN/True Negative: This when a case is truly negative and the resulting prediction is negative [10].
2) TP/True Positive: This when a case is positive and prediction result is positive [10].
3) FN/False Negative: This when a case is positive but the predicted result is negative [10].
4) FP/False Positive: This when a case is negative but the prediction is positive [24]. A false positive is not desirable at all as this can be very misleading result [10].

We calculate precision by using the formula: precision = TP/(TP + FP) [10] We refer to figure 18 for random forest precision, for support vector algorithm its figure 19 and lastly for naive bayes we have figure 20.
2) RECALL
The recall matrix looks into how many positive cases the classifier was able to catch. With the recall matrices we identify the ability of the classifier to find all positive instances [10]. We calculate recall by using the formula: recall = TP / (TP + FN) [10] We refer to figure 18 for random forest recall, for support vector algorithm its figure 19 and lastly for naive bayes we have figure 20.

3) F1 SCORE
F1 score is the “weighted harmonic mean of precision and recall” [10]. Where with the f1 score the best score is 1.0 and the worst score which could be recorded is 0.0 and lastly based on the rule of thumb “the weighted average of f1 should be used to compare classifiers, not global accuracy.” [10]
We calculate f1 score by using the formula: F1 score = 2 * (Recall * Precision) / (Recall + Precision) We refer to figure 18 for random forest F1 score, for support vector algorithm its figure 19 and lastly for naive bayes we have figure 20.

C. CONFUSION MATRIX
Pearson’s correlation coefficient is utilised in the research to find a statistical relationship or association between two variables in the dataset, Pearson’s correlation coefficient is based on a method of coefficient [11]. In this part of the document we look at the different symptoms that exists to find if they have either is a positive or negative correlation between these variables. We look at figure 21 showing a Seaborn heat map from Pearson’s Coefficient. The heat map show a relationship between two symptoms. The test set was used for the heat map diagram. The heat map is made of the first 20 out of 133 symptoms of the dataset. Most of the symptoms do not show any strong positive correlation nor do they show a negative correlation. In the table below, table 6 we see different variables with moderate and strong positive correlation with each other. This shows that there is a relationship between shivering and continuous sneezing. Diseases such as influenza have shivering, chills and continuous sneezing as symptoms, [19]. This explains why shivering and continuous sneezing has a strong positive correlation.

D. PROTOTYPE SYSTEM
1) USER INTERFACE
Figure 22 shows a snip of the prototype system predicting diarrhoea and saving all the data on a database created under PostgreSQL. The system makes use of three different tables where the first table is Patient_info, the next tale is Patient_symptoms which stores the symptoms of patients together with the national Id as the foreign key and the last table which is disease_prediction which store the predicted disease together with the national Id of the patient, See figure 12 for the tables in which the data will be stored.
TABLE 6. Variables with high correlation based on the support vector machine.

| Symptom 1       | Symptom 2      | Correlation Score | Positive or Negative Correlation |
|-----------------|----------------|-------------------|----------------------------------|
| weight gain     | cold hands and feet | 1.0               | Positive Correlation             |
| mood swings     | weight gain     | 0.7               | Positive Correlation             |
| mood swings     | cold hands and feet | 0.7               | Positive Correlation             |
| Spitting urination | Burning micturition | 0.7               | Positive Correlation             |
| Shivering       | Continuous Sneezing | 0.7               | Positive Correlation             |
| Ulcer on Tongue | Acidity         | 0.7               | Positive Correlation             |
| Spitting urination | Stomach pain   | 0.56              | Positive Correlation             |
| Ulcer on Tongue | Stomach pain    | 0.56              | Positive Correlation             |
| Chills          | Continuous Sneezing | 0.5               | Positive Correlation             |

For the text such as the national identity of the patient it is made of 13 digits and the text box should only take 13 digits as it will be restricted. This is done to make sure that the information entered is correctly. National identity number is very important in the whole system.

The prototype system based on tkinter, see figure 22 has different buttons for it function. We have four major buttons which are save patient information, save symptoms predict and about. All these button play a major role where save patient information saves the data in table patient_info figure 12. The button save symptoms stores information from all the different symptoms selected and the national identity document of the individual on table patient_symptom, figure 12. The about page simply gives more information about diarrhoea and further directs individuals to a medical website which has more information about diarrhoea, so that the population can be well informed.

2) DATABASE

The use of structured query language is vital for the system to effectively retrieve cases of diarrhoea of children under the age of 5 and where these cases are within the country. In this study postgresql was used as the database to store the data that is collected and health officers will make use of the data to further keep surveillance of the diarrhoea situation and come up with and effective contingency plan if there were an outbreak which would occur. I created a database and named it DiarrhoeaDB and it is made of or consists of three major tables, these are patient_info, patient_symptom and disease_prediction, an E-R diagram is seen in figure 12.

The purpose of the thesis are views to simply view the desired data see figure 23 and count to simply count the number of diarrhoea cases under the age of 5, see figure 26 for results. The queried data will allow health workers to see how many people had a diarrhoea prediction based on the symptoms given by the patient. The information stored in the database is then further imported to a csv file to allow further plotting in google maps.

E. LOCATION

The location section is the last part of the thesis where we now locate the places based on the input data from the converted csv file. With the data from the database converted and view simply set to download data where the predicted prognosis is diarrhoea. This will allow a map creation for just people who have diarrhoea symptoms and are under the ages of five. The different pins will be dropped on the map and when you view place your cursor on the pin, information that will be selected to be part of the pin will be viewed. This will allow tracking and deeper surveillance.

With my maps you are able to perform different functions such as finding the distance in which the cases are located and this provides further information which can be used in surveillance. Both maps and postgresql are vital in the response and surveillance of diarrhoea in order to make sure that there is no outbreak which occurs. This creates vigilance within the surveillance of diarrhoea.

We mentioned in the previous chapter that for this research, out of the 4 regions in Eswatini one region will be looked at which, is the Manzini region and of the 18 constituencies, 4 will be looked at which are Lamgabhi, Mhlambanyatsi, Lobamba Lomdzala and Ntondozi. In figure 24, we see different points on the map. These different plots show where the people were from. The plot is based on the constituency and region provided by the patient. The information is imported to my maps using a csv file downloaded from the DiarrhoeaDB created in postgresql. Figure 25, displays a sample of dummy information of a patient used to test the database.
The tables display the age of the patient, their cell number, the prognosis and the symptoms they experienced. One feature about my maps you can choose which type of data you want to view and this provides security for information of the patient.

Figure 25 shows a different region and displays the similar information about the patient, it shows more of the patient information. Which information to display can be chosen. For the purpose of this study, the focus is on the Manzini Region and its constituencies as it is the largest region in Eswatini and has the highest population of people. The research will further look into 5 of the 16 constituencies in Eswatini.

V. DISCUSSION AND CONCLUSION
Diarrhoea may look simply as just passing loose stool for a period of time and it may seem normal, not very serious. However, it is a very serious disease and it is amongst the top 10 diseases that lead to death of children under the age of 5. Developing and designing ways to detect diarrhoea at an early stage before it spreads is vital because, as it can be transferred from one person to the other. There are factors which contribute to diarrhoea, these factors include: the lack of access to clean water which are used to drink, cook and water gardens. Another contributing factor is the lack of clean lavatories for people, to use mainly in the rural areas.

It is important that a continuous surveillance of diarrhoea is done from the time when there is little evidence of an outbreak, to during an outbreak and after the outbreak. This continued surveillance is important as to curb the probability of an outbreak. The most common diarrhoea in The Kingdom of Eswatini is acute Diarrhoea. With the help of machine learning, an algorithm can be trained using symptoms of different diseases and can be trained on what the prognosis is for each of these diseases.

With the use of a database and SQL commands, it easier to find and locate people whom are under the age of 5 and have experienced diarrhoea symptoms. A record of where they reside and their identification number is stored and used to pin their location on my maps. This will assist with tracking and tracing of where the diarrhoea cases are and how many people in the general region have had diarrhoea. For the purpose of this study, the focus is on the Manzini Region and its constituencies as it is the largest region in Eswatini and has the highest population of people. The research will further look into 5 of the 16 constituencies in Eswatini.

With the 133 symptoms and 42 prognosis the dataset contained, the prototype system is able to effectively predict diarrhoea. The population which is mostly affected by diarrhoea are children under the age of 5. Saving the information in a database and supplying the information to the main health database system will create a trail of information on which constituencies within the four regions is prone to diarrhoea and having the covariates playing a major role in the spread of diarrhoea. The information will allow further surveillance and further action by the health department.

The supervised machine learning algorithms utilized in this research are the Support Vector Machine, Naive Bayes and the Random Forest Classifier. Each of these algorithms were first used in the prototype system to check if they accurately predict diarrhoea. In which only the support vector machine correctly predicted diarrhoea. The two algorithms, naive bayes and random forest classifier achieved an average accuracy of 97.62% When given the six symptoms. Support vector machine achieved an average accuracy of 100%.
The health department is the most important department. Creating more solutions with the aim at gathering more information from data collected is vital. Getting symptoms from ten different people and storing that information can bring more knowledge to the health department. With that knowledge extracted from the algorithms, one can find evidence to assist in making decisions and taking action about a potential outbreak of a disease before it causes havoc is important and necessary.

In conclusion, with the current pandemic, the health department and its facilities have been overemphasized. The departments need further assistance. Machine learning algorithms have been used in many other disease prediction cases such as heart attack and kidney disease. With the help of storing patient information, this can help better build and develop algorithms which can be highly efficient in the medical field. This paper further contributes to the creation of a data footprint of symptoms most people experience and the predicted diagnosis will also be stored. Having this information and more personal information about patient will allow further analysis in the future, bettering the health care in Eswatini.

A. LIMITATIONS

- Availability of a larger dataset for symptoms and prognosis. This problem then further channels which model to pick as other models would cause an underfitting problem. Experiments were conducted using long short-term memory loss neural network but then an underfitting problem occurred. Hence I had to work with machine learning algorithms which will be able to train on a small dataset. This then limited me or channelled me to choose certain algorithms. My choice of algorithms was narrowed down.

- Availability of a variety of diseases containing different symptoms

- Availability of current statistic on how Eswatini is doing with diarrhoea cases.

- Statistics of the availability of clean water within the different constituencies in the different regions and also the number of households with limited access to clean or lavatories as these two are major covariates of diarrhoea.

REFERENCES

[1] S. Uddin, A. Khan, M. Hossain, and M. Moni, “Comparing different supervised machine learning algorithms for disease prediction,” BMC Med. Inform. Decision Making, vol. 19, no. 1, pp. 1–16, 2019, doi: 10.1186/s12911-019-1004-8.

[2] Ministry of Education, Eswatini Diarrhoea Contingency Plan, 2018.

[3] Y. Deepthi, P. Kalyan, M. Vyasa, K. Radhika, K. Babu, and V. Krishna, Disease Prediction Based on Symptoms Using Machine Learning, vol. 664. Singapore: Springer, 2020. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3861426

[4] N. E. Soboka, “Associations between improved water supply and sanitation usage and childhood diarrhea in Ethiopia: An analysis of the 2016 demographic and health survey,” Environ. Health Insights, vol. 15, Jan. 2021, Art. no. 117863022100255, doi: 10.1177/117863022100255.22

[5] S. Radhika, S. Ramya, V. Rukhmani, and A. Ranjitha. (2020). Disease Prediction Based on Symptoms Using Machine Learning. [Online]. Available: https://ejmcm.com/article1944.html

[6] Finding Associated Concepts With Text Analysis. Accessed: Nov. 13, 2020. [Online]. Available: http://www.nactem.ac.uk/facta/index.html

[7] World Health Organization. (2017). Diarrhoeal Disease. Accessed: Oct. 16, 2020. [Online]. Available: http://www.who.int/news-room/factsheets/detail/diarrhoeal-disease

[8] S. Radhika, S. Ramya, V. Rukhmani, and A. Ranjitha. (2020). Disease Prediction Based on Symptoms Using Machine Learning. [Online]. Available: https://ejmcm.com/article1944.html

[9] G. Maphalala, N. Phungwayo, G. Masona, N. Lukhele, and G. Tsegaye. (2018). Early Impact of Rotavirus Vaccine in Under 5year Old Children Hospitalized Due to Diarrhea, Swaziland. Vaccine. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/28778615

[10] T. Muthukrishnan. (2018). Understanding the Classification Report Through Sklearn. Accessed: Oct. 31, 2021.

[11] P. Schober, C. Boer, L. Schwarte. (2018). Correlation Coefficients: Appropriate Use and Interpretation. Anesthesia Analgesia. [Online]. Available: underlinemephimpress://https://www.researchgate.net/publication/323388613_Correlation_Coefficients_Appropriate_Use_and_Interaction

[12] S. Grammurohit and C. Sagarnal. “Disease prediction using machine learning algorithms,” in Proc. Int. Conf. Emerg. Technol. (INCET), Jun. 2020, pp. 1–7. [Online]. Available: https://ieeexplore.ieee.org/document/9154130

[13] H. Chauhan, N. Naik, A. Halpati, J. Patel, and A. Prajapati. (2020). Disease Prediction using Machine Learning. [Online]. Available: https://www.irjet.net/archives/V7/I5IRJET-V7I5385.pdf

[14] Dataset. Accessed: Nov. 9, 2020. [Online]. Available: https://www.kaggle.com/itachi9604/disease-symptom-description-dataset?select=Symptom-severity.csv

[15] Y. Zoabi, S. Deri-Rozov, and N. Shomron. (2021). Machine Learning-Based Prediction of COVID-19 Diagnosis Based on Symptoms. [Online]. Available: https://www.nature.com/articles/s41746-020-00372-6

[16] G. Maphalala, “Prevalence of rotavirus in under 5 year old children hospitalized for acute gastroenteritis, swaziland, 2013–2014,” Sci. J. Public Health, vol. 5, no. 4, p. 353, 2017. [Online]. Available: http://www.sciencepublishinggroup.com/journal/paperinfo?journalid=251&doi=10.1164j.sjph.20170504.22

[17] P. Miló, B. Mhlanga, and A. Senzanje, “Emerging issues in water resources management in Swaziland,” Tech. Rep., 2020.

[18] E. Mousseme and F. Cedeño. (2020). Machine-Learning Forecasting for Dengue Epidemics—Comparing LSTM, Random Forest and Lasso Regres-Sion. [Online]. Available: https://www.medrxiv.org/content/10.1101/2020.01.23.20018556v1

[19] S. Ghebrehewet, P. MacPherson, and A. Ho. (2016). Influenza. BMJ (Clinical Research Ed.), 355, i6259. [Online]. Available: https://www.bmj.com/content/355/bmj.i6259

[20] B. Kanche, Sud K. Mahale, (2016). Study of Machine Learning Algorithms for Special Disease Prediction Using Principal of Component Analysis. [Online]. Available: https://ieeexplore.ieee.org/document/7955260

[21] R. Abubakar and O. Olatunji, “Computational intelligence-based,” Tech. Rep., 2020.

[22] A. Ogbo, K. Agho, P. Ogeloka, S. Wolffenfeld, A. Page, and J. Eastwood. (2017). Prevalence and Modifying Effect of Sanitation on Infant and Young Child Feeding Indicators Associated With Diarrhoea in Sub-Saharan African Countries, With High Burden of Diarrhoea Mortality. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0171172

[23] D. Harini and M. Nathes. (2018). Prediction of Probability of Dis-Ease Based on Symptoms Using Machine Learning Algorithm. [Online]. Available: https://www.irjet.net/archives/V5/5/IRJET-V5I574.pdf

[24] T. Joachims. (2002). Learning to Classify Text Using Support Vector Machines: Methods, Theory, and Algorithms. [Online]. Available: https://www.researchgate.net/publication/257131900_Learning_to_Classify_Text_Using_Support_Vector_Machines_Methods_Theory_and_Algorithms

[25] F. Marouane. (2020). Disease Prediction Using Machine Learning. [Online]. Available: https://www.researchgate.net/publication/347381005_Disease_Prediction_U sing_Machine_Learning

[26] C. Aishwarya, K. Suvarchala, B. Aravind, G. Shashank, M. Anand, and V. Krishna, “Prediction of disease using machine learning and deep learning,” in Energy Systems, Drives and Automations, 2020, pp. 69–79, doi: 10.1007/978-981-15-5089-8_7.

[27] A. Depeursinge (2017). Biomedical Texture Analysis. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128121337000028
SIBONGAKONKE KWANELE ZUNGU received the Bachelor of Science degree from the University of Swaziland, where she did a double major in computer science and geography, planning and environmental science. Her research interest includes creating digital data footprint for her country Eswatini and at the same time developing algorithms to help prevent preventable circumstances through big data analysis.

QI-XIAN HUANG is currently pursuing the Ph.D. degree with the National Tsing Hua University, Hsinchu, Taiwan. His research interests include deep learning for computer vision, beyond 5G networks, and the Internet of Things security.

MIN-YI CHIU is currently pursuing the Ph.D. degree with the Institute of Information Systems and Applications, National Tsing Hua University, Taiwan. His research interests include information security and blockchain applications, system design, network security, and forensics.

HUNG-MIN SUN received the Ph.D. degree in computer science and information engineering from National Chiao-Tung University, Hsinchu, in 1995. He was an Associate Professor with the Department of Information Management, Chaoyang University of Technology, from 1995 to 1999, the Department of Computer Science and Information Engineering, National Cheng-Kung University, from 2000 to 2002, and the Department of Computer Science, National Tsing Hua University, Hsinchu, from 2002 to 2008. He is currently working as a Full Professor with the Department of Computer Science, National Tsing Hua University. He has published more than 200 international journal articles and conference papers. His research interests include network security, cryptography, blockchain, and automatic trading. He was the program committee member of many international conferences. He serves as the editor member for many international journals. He received many best paper awards in academic journal and conferences, including the Annual Best Paper Award from the Journal of Information Science and Engineering in 2003, the Best Paper Award in MobiSys 2009, NSC 2005, NISC 2006, NISC 2007, CISC 2009, ICS 2010, ICMS 2019, and IEEE ICKII 2019. He also received the Y. Z. Hsu Scientific Paper Award from the Far Eastern Y. Z. Hsu Science and Technology Memorial Foundation in 2010, the Outstanding Research Award from World Congress on Information Technology Applications and Services in 2015, and the Outstanding Professor in electrical engineering from the Chinese Institute of Electrical Engineering in 2014. He was the Program Co-Chair of the 2001 National Information Security Conference. He was the Honor Chair of the 2009 International Conference on Computer and Automation Engineering, the 2009 International Conference on Computer Research and Development, and the 2009 International Conference on Telecom Technology and Applications. He is also the General Chair of ACM AsiaCCS in 2020.

* * *

VOLUME 9, 2021 152959