Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 04, April-2023

Prediction of Waterborne Diseases using Machine Learning Tools

Asmita Patil
Department Information Technology
Government College of Engineering
Karad, India

Sanjeev Wagh
Department Information Technology
Government College of Engineering
Karad, India

***Abstract*—Worldwide, waterborne illnesses continue to be the main contributors to human illness and mortality. Waterborne illness is similar to illnesses brought on by food contamination and can be brought on by microorganisms often linked to the food-borne transmission. breakouts connected to meals or occasions where food was initially considered to be a factor can be brought on by a waterborne illness occasionally. You can prevent more than 95% of these. As the primary preventive tool for preventing chronic diseases, ensuring that everyone has access to water and sanitation is one of the Sustainable Development Goals the United Nations has set for 2030. Even though the basics of sanitation and water treatment are well known, billions of people have been denied access due to a lack of capital assets, visionary leadership, and logical prioritization, Emerging infections resistant to traditional water purification, chemical pollutants, recognizing both endemic and pandemic waterborne diseases, and comprehending connections to the environment are some of the challenges. Satellite photography and new mathematical techniques are shedding new light on the study of aquatic infections. As the diagnosis of waterborne disease is expensive, using machine learning algorithms prediction of waterborne diseases is more accurate and easily. The algorithms utilized in this project are naive Bayes, decision trees, and random forests.**

Keywords*— *Waterborne; Foodborne; Contamination; Microorganisms; Sanitation;Machine Learning

I. INTRODUCTION

Ailments caused by microscopic organisms, such as viruses and bacteria, that are ingested through contaminated water or by coming into contact with feces are referred to as "waterborne illnesses." If everyone had access to clean water, appropriate sanitation, and hygiene practices, these diseases would not exist. Governments, non-governmental organizations, and local communities have achieved great strides in the fight against waterborne infections over the past 20 years. There is still a lot of work to do.

Most healthy people who are exposed to water contaminated with these pathogens do not become ill. A person who is older than 50, a smoker now or in the past, has a lung ailment or has a compromised immune system are some of the groups of people who are more susceptible to illness.

It is easy to prevent cholera when travelling since it is a watery illness. Avoid eating raw fish (no sushi), wash your hands frequently, and Eat only products that you can peel yourself, such as oranges, bananas, and avocados. Consume a lot of fresh water. If there are no handwashing stations available, cholera can spread throughout an entire village. According to studies, 40% of households in underdeveloped countries like Ethiopia lack the supplies—safe water, soap, and a bathroom—needed to properly wash their hands. For these groups, keeping a clean environment and avoiding illness is nearly impossible. In remote villages, Life Water teaches families how to construct their handwashing equipment, assisting in the fight against cholera.

Dysentery Extreme diarrhea and the presence of blood or mucus in the stool are symptoms of the waterborne disease dysentery, which is brought on by an intestinal infection. Dysentery is a solid reason to frequently wash your hands because the disease is largely spread by people who don't practice good hygiene. It can be caused by bacteria, viruses, or parasites as well as tainted food, water, or feces. If patients with dysentery do not promptly replenish lost fluids, their lives may be in jeopardy.

Smoking fever Typhoid fever is common in underdeveloped regions of developing countries, despite being uncommon in wealthy nations. Up to 20 million people are thought to contract the sickness annually around the globe. It is very contagious and spreads through tainted food, unclean water, and poor hygiene.

When left untreated, cholera, an acute stomach illness that causes vomiting and watery diarrhea, can swiftly cause severe dehydration and even death. According to the World Health Organization (WHO), cholera affects 1 million to 4 million people worldwide and claims up to 143 000 lives per year. Instances were reported in 54% of African nations in 2016. Death rates decrease as people get access to wholesome food and water.

Waterborne infections have become a major concern for global health, and it is crucial to quickly and accurately detect them. So today it's crucial to discover these diseases early and avoid them.

While the most frequently reported symptoms of a waterborne sickness are diarrhoea and vomiting, other symptoms can include skin, ear, respiratory, or eye issues. A community's lack of access to clean water, adequate sanitation, and good hygiene (WASH) are key contributors to the spread of waterborne illnesses.

Most often, transmission occurs when a person consumes contaminated food, water, or feces. Aerosols created during frequent vomiting have been linked to viral infection epidemics via aerosol transmission.

The most noticeable symptom is severe diarrhoea, which is accompanied by nausea, vomiting, and stomach discomfort. However, the dehydration that results from this

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 04, April-2023

may cause electrolyte abnormalities, abrupt renal failure, and encephalopathy. Rarely, protracted or more severe complications from waterborne disease, include anemia, shock, hemolytic uremic syndrome, spontaneous abortion, convulsions, and liver, heart, or lung disease.

Waterborne illnesses other than diarrhoea include listeriosis, hepatitis A, etc. The involvement of the entire body is evident in these disorders.

Symptoms of the more typical viral, bacterial, and parasitic infections include nausea, vomiting, diarrhoea, headaches, fevers, and kidney failure. Additionally, infectious disorders like hepatitis can manifest.

To detect disease early, for our experiment, by using symptoms disease detector kit was created. The foundation for using a machine learning algorithm for waterborne disease detection is briefly explored. Future research and conclusions are presented in the document.

II. LITERATURE SURVEY

Case Fatality Rate, which is the death rate as a result of unusual circumstances Diarrhoea outbreaks in Indonesia continue to exceed government expectations [1]. Several variables might lead to diarrhea, but the one that has the biggest impact on symptoms is the Diarrhoea case fatality rate is unknown. Consequently, this study's goal is to develop data patterns in the form of classification from data on diarrhea outbreaks Diarrheal Case Fatality Rate can be determined using a categorization rule. Classification employed the Naive Bayes algorithm and the C4.5 algorithm. The popular algorithm is the C4.5 algorithm. While Naive Bayes method is well-liked with probabilistic decision trees, a classification strategy. Utilizing the phases of Knowledge Discovery in research in the database after acquiring the categorization rule, the Confusion Matrix evaluated the rule, as well as the Receiver Operating Characteristic Curve. The assessment was carried out utilizing training data and data testing. The analysis's findings reveal that the C4.5 algorithm has a higher greater accuracy than Naive Bayes. While the primary determinants of Case Fatality Shelter and sanitation contribute to a rise in diarrheal illness rates.

The advancement of big data in the healthcare sectors and biomedical has made accurate analysis of medical data possible [2]. enhances patient care, early disease detection, and social services. when medical data of high quality is lacking, the study's accuracy suffers. Additionally, several places display distinctive manifestations of particular local diseases, which could decrease the ability to anticipate disease outbreaks. The system offers machine-learning techniques for accurate disease prediction across a range of conditions. instances in societies where sickness is common. It tests the modified estimation models over actual hospital data. data gathered. The employment of a latent factor model to reconstruct the data helps to get around the problem of incomplete data. lacking data It is an experiment on a localized, persistent brain infarction. Utilizing organized and With the Machine Learning Decision Tree technique and Map Reduce, unstructured data from hospitals algorithm. To the best of knowledge, no work has been done in the field of medical big data analytics. focused on both forms of data. When compared to some common estimation algorithms, the computation accuracy of author suggested technique achieves 94.8% with a faster convergence rate than the CNN- based risk prediction for unimodal diseases

People today deal with a variety of ailments. due to their way of life and the state of the climate. While early illness detection becomes a crucial task [3]. But the doctor finds it too challenging to precisely ascertain based on symptoms. Data Mining is crucial for forecasting the disease. in solving this issue. There is a medical science yearly rise in data that is significant. Author suggested general disease prediction depending on the patient's symptoms. Author utilize technology for disease prediction. Support Vector Machine (SVM) learning technique for accurate disease prediction. illness data set Symptoms are necessary for disease prognosis.

Africa is where 54% of the world's sickness is found. the burden brought on by the inability to get clean drinking water, with the majority of endemic zones or rural region people getting access to water from potentially dangerous public water faucets [4]. However, the high-cost laboratory procedures and resources used to identify water-borne illnesses in water treatment facilities Like cholera cannot be widely spread through all of those faucets, ensuring everyone has access to safe water, whenever and wherever. Thanks to the fusion of artificial intelligence with the internet of things (IoT) The prediction of water bone disease cholera can be made using artificial intelligence (AI). done by observing the physicochemical patterns of water. However, Modern Internet of Things/ artificial intelligence solutions relies on a cloud-centric infrastructure. framework transferring collected data from cutting-edge water parameter sensors cloud storage for data.

The prognosis of numerous water-borne illnesses including cholera and typhoid forms the basis of the paper [5]. To do this, authors employ temperature and pH sensors This will enable authors to prevent the loss of many lives to water-borne illnesses. Authors are using the suggestions from the sensor and transmitting the data to the technology, which analyzes the water quality and forecasts the proportion of persons who will be impacted by various uses with the aid of machine learning algorithms, to diagnose disorders.

In India's most southern region, dengue is a fever spread by mosquitoes [6]. It results from females. The primary signs and symptoms of dengue are fever, bleeding, and pain behind the eyes. To save the patient's life, early detection of symptoms such as abdominal pain, exhaustion, and appetite loss is crucial. human from this fatal illness. Utilizing classification algorithms enables early illness prediction. In this study, the Bayes belief network classification method is utilized to forecast the likelihood of various estimating the risk of an illness occurring.

Despite various efforts by the government at all levels and other entities concerned with water and its safety, waterborne illnesses remain a serious public health and environmental problem [7]. Spending so much money on water research was worthwhile, but the intended outcomes have not been achieved because waterborne diseases continue to be a problem in developing countries, with Africa and Asia being the most affected. The primary causes of the growing frequency of waterborne infections are without a doubt the

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 04, April-2023

lack of pipe-borne water and rural populations' reliance on surface waters that are regularly contaminated with feces. These rural communities' lack of access to clean water and shoddy hygiene practices are a major concern because they greatly contribute to the spread of illnesses that are contracted through contact with water. In addition, improper environmental practices that encourage the development of insects and other vectors in residential areas contribute to the increase in the occurrence of waterborne illnesses. This paper focuses on the many methods employed in the classification of waterborne diseases and their modes of transmission, as well as the bacteriological analysis of water.

III. METHODOLOGY

The three data mining algorithms—a Naive Bayes classifier, a decision tree classifier, and a random classifier—are used to create the illness prediction system. The algorithm's description and operation are provided below.

A. Decision Tree Classifier

The decision tree method produces classification models with a tree-like topology. It divides the information into ever-smaller subsets and predicts a target value (illness) by learning a sequence of explicit if-then rules on feature values (in our instance, symptoms). Decision nodes and leaf nodes make up a decision tree.

- Decision node: Contains at least two branches. All of the symptoms in the study we've provided are viewed as decision nodes.
- Leaf node: This node symbolises the classification, or choice, of any branch. In this instance, the leaf nodes represent the diseases.

J.R. Quinlan's ID3 method is one of the fundamental algorithms we have employed in our study. Using a top-down, greedy search over the provided columns, ID3 chooses the attribute (symptom) that is most effective for categorising a particular set after testing each column (attribute=symptoms) at each node. ID3 employs information gain and entropy to determine which symptom is the best candidate for a decision tree.

- Entropy E(C), using the frequency table of one attribute, where C is the current state (outcomes that have already occurred) and P(h) is the probability that event h will occur given that condition C.
- The entropy E (C, A) is calculated using the frequency table of the two attributes C and A, where C is the present state with attribute A and A is the considered attribute, and P (h) is the probability of an event H of attribute A.

Information Gain: Following the finalization of an attribute A, information gain, also known as Kullback-Leibler divergence, is denoted by the symbol IG (C, A) for a state C.

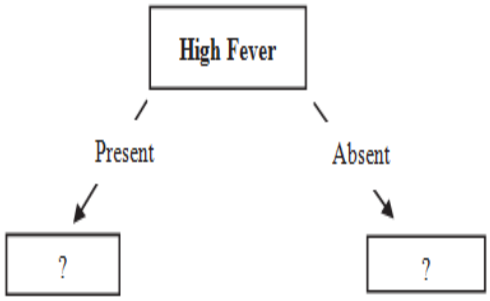

Fig. 1. Decision Tree Flow Chart

There are three additional symptoms left after using High fever: Vomiting, Shivering, and Muscle Wasting. Additionally, the sub-trees for High fever have two alternative values: Present and Absent. Sub-tree of the Present beginning:

Four of the seven situations where the attribute value of high fever is present result in dengue, and three in malaria.

B. Random forest Classifier:

Random forest is a versatile, user-friendly machine learning method that, in the majority of cases, produces excellent results even without hyper-tuning. Overfitting is a major drawback of the decision tree method, as noted in the decision tree. It seems that the tree has retained the information.

This issue is prevented with Random Forest: It is a type of group learning. Using different algorithms or the same method repeatedly is referred to as ensemble learning. A group of decision trees called a random forest. In Random Forest, the more of these decision trees there are, the more accurate the generalization.

C. Naive Bayes Classifier:

The core tenet of the Naive Bayes classifier is that each feature contributes equally and independently to the final result. Due to the fact that it uses less processing resources, it has the advantage of operating quickly even on huge datasets.

Algorithms

A. Decision Tree

- As the name says all about it, it is a tree which helps us by assisting us in decision-making.Used for both classification and regression, it is a very basic and important predictive learning algorithm.

It is different from others because it works intuitively i.e., taking decisions one-by-one.

Non-parametric: Fast and efficient Decision tree considers the most important variable using some fancy criterion and splits dataset based on it. It is done to reach a stage where we have homogenous subsets that are giving predictions with utmost surety.

B. Random forest

-Random Forest is a very popular ensemble learning method which builds a number of classifiers on the training data and combines all their outputs to make the best predictions on the test data. Thus, the Random Forests algorithm is a variance minimizing algorithm that uses randomness when making split decision to help avoid overfitting on the training data.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 04, April-2023

C. Naïve Bayes

-The Naive Bayes Algorithm is one of the crucial algorithms in machine learning that helps with classification problems. It is derived from Bayes' probability theory and is used for text classification, where you train high-dimensional datasets.

IV. IMPLEMENTATION AND RESULT

We suggested a system for predicting waterborne diseases based on a machine learning algorithm. To categorise the data, we used machine learning methods. Accuracy is increased when applying machine learning algorithms to forecast diseases and all sub-diseases. processing data to make precise, test result-based predictions about waterborne disease. We were able to produce an exact prediction of a waterborne disease by providing the input. With the use of this method, it was possible to anticipate waterborne diseases quickly and affordably.

Algorithm performance on data: The system was trained on the medical records of 4920 patients who were predisposed to 42 diseases as a result of the comorbidity of different symptoms. To prevent overfitting, 95 out of 132 symptoms have been taken into account.

We tested the effectiveness of each method on the dataset using the K fold cross validation technique (K=5).

Algorithm's accuracy score was:

TABLE I. THE ACCURACY TABLE

Algorithm used	**Accuracy score**
Naïve Bayes	89.9224806
Decision Tree	87.5968992
Random Forest	68.9922480

From the above table, we can infer that each algorithm has a different accuracy score. the accuracy in terms of percentage: 89.92%, 87.59%, 68.99%.

V. CONCLUSION

The recommended system has shown that the Bayesian belief network classification technique with a probability distribution table will help in predicting the likelihood of a variety of qualities to the presence of the associated illness. This research gives a thorough comparison of three algorithms' performance on a medical record, with the naive bayes algorithm producing an accuracy of up to 89%, the Decision tree classifier producing an accuracy of up to 87% and the Random Forest classifier producing an accuracy of up to 68%. It is recommended to use more advanced techniques in the future to increase the accuracy.

REFERENCES

[1] Wahyudi, Mochamad, and Anik Andriani. "Application of C4. 5 and Naïve Bayes Algorithm for Detection of Potential Increased Case Fatality Rate Diarrhea." *Journal of Physics: Conference Series*. Vol. 1830. No. 1. IOP Publishing, 2021.

[2] Vinitha, S., et al. "Disease prediction using machine learning over big data." *Computer Science & Engineering: An International Journal (CSEIJ)* 8.1 (2018): 1-8.

[3] Kaur, Sandeep, and Kuljit Kaur Chahal. "Hybrid ANFIS-genetic algorithm based forecasting model for predicting Cholera-waterborne disease." *International Journal of Intelligent Engineering Informatics* 8.4 (2020): 374-393.

[4] Shinde, Revati, et al. "Disease Prediction Using Machine Learning." (2021).

[5] Ogore, Marvin Muyonga, Kizito Nkurikiyeyezu, and Jimmy Nsenga. "Offline Prediction of Cholera in Rural Communal Tap Waters Using Edge AI inference." *2021 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2021.

[6] Shivam Kumar, et al."" water disease prediction device". IJCRT 2021.

[7] Anitha, A., and S. Freeda Jebamalar. "Predicting Dengue Using Bayes Net Classifier." (2018).

[8] Kurtah, Pratima, Yusrah Takun, and Leckraj Nagowah. "Disease propagation prediction using machine learning for crowdsourcing mobile applications." *2019 7th International Conference on Information and Communication Technology (ICoICT)*. IEEE, 2019.

[9] Grampurohit, Sneha, and Chetan Sagarnal. "Disease prediction using machine learning algorithms." *2020 International Conference for Emerging Technology (INCET)*. IEEE, 2020.

[10] S. Maharditya Restu ,"Dengue hemorrhagic fever(DHF)classification for patient in puskesmas usingnaïvebayes algorithm".