A Framework Questionnaire for Diagnosing Infectious Disease Using Machine Learning Techniques

Hayder Hussein Thary, Khamis A Zidan

Computer Engineering Department, College of Engineering, Al-Iraqia University, Baghdad, Iraq
E-mail: hayder_thary@aliraqia.edu.iq

Abstract. Infectious diseases weigh down the communities in the world and scientists to spend more effort via keeping tracking of evolving treatment and detecting methods. These diseases may lead to harm life of people. Early diagnosis could significantly support healthcare specialists to save more lives. Additionally, the pandemic leads to maximizing hospitals visits, medical clinics and healthcare centres. The international health organizations also have shown that there has been a rapid growth of infected cases. Therefore, correct diagnosis has become a pressing problem. Consequently, automated diagnosis becomes a sensible solution to the problem of these diagnosis challenges. This study was conducted to identify the most common infectious diseases in the Iraqi society using a well-designed questionnaire and a proposed automated diagnostic technique. Firstly, the top diseases questionnaire is distributed around the city of Baghdad to different medical clinics. The results from the preliminary analysis of the collected responses (115 responses) showed that the most common widespread diseases in the Iraqi community are diabetes, flu, and typhoid. This was followed by another questionnaire for the identification of symptoms and blood test variables for these diseases. It is worth pointing out that there are not sufficient and updated studies dealing with the diseases that attack the Iraqi community. Toward the automated diagnosis, both infectious diseases (flu and typhoid), identified symptoms are employed as feature space with one of the machine learning techniques. For the results evaluation different measures, such as accuracy, confusion matrix, and efficient verification via ROC, have been used to indicate the system performance. The result shows that typhoid disease has significant diagnosis accuracy of 98% compared to the others. While three machine learning systems named (Native Bayes, Linear discriminant, and Ensemble (subspace discriminant)) were used to diagnose flu disease. The resulting accuracy of all three models are 92% which shows good performing. Therefore, the proposed method shows precise accuracy and systematic manner for analyzing infectious diseases.

Keywords. Infectious diseases, Questionnaire, Flu, typhoid, Diabetes, Symptoms, Machine learning.

1. Introduction
Infectious diseases thrive in conflict settings. Therefore, outbreaks can spread rapidly in communities and populations because of deteriorating nutrition, overcrowding of population in small cities and unsanitary conditions in displaced camps, and deteriorating health care infrastructure [1]. Hence, infectious diseases have recently become a significant cause of mortality. In 2016, infectious diseases caused 3.0 million deaths worldwide [2]. In epidemiology, one aims to investigate the process of identifying and mapping emerging diseases in a specific population. In order to control related health
problems and improving population health [3]. The main goal for epidemiologists is to detect disease outbreaks early and minimize the costs of healthcare [4]. These diseases are the problems caused by organisms and can be transmitted or spread from one person to another [5]. Though, these diseases are the primary cause of any given population illness [6]. Infectious agent-causing diseases include various micro-organisms including bacteria, viruses, fungi, protozoa, and helminths. A new class of infectious agents, the prions, has also been recognized recently [7]. The intensity and volume of population movements activity of the Iraqi population have changed radically, increasing the population risk of infectious diseases, especially tuberculosis, hepatitis, allergic conjunctivitis, and German measles [8,9]. Biotechnology has advanced a great deal lately. As a result, biologists enter the big data. With the emergence of high-performance resulting incessantly in a simple output of data, the science of applied biology is being incorporated into the field of big data [10,11]. Biotechnology data collection and review machines are installed in new and existing hospitals to allow them to gather and exchange data in large information systems. Machine Learning Systems are very useful in processing medical data and a great deal of work is being performed on diagnosis issues [12]. Artificial Intelligence (AI) and machine learning techniques can have a wide variety of applications. These frameworks, in particular, have a role in healthcare, especially in medical diagnosis. A medical diagnosis is a method of classification. A physicist needs to examine a series of signs and symptoms before he or she diagnoses an illness that makes his or her work complicated [13–15]. Accordingly, the problem with modern medicine, however, is the correct diagnosis of diseases which is before medication [16]. By using the questionnaire method which is a research tool that contains a sequence of questions and their prompts to gather information and opinions of the respondents. Sir Francis Galton was the first person who invented the questionnaire, who is a British anthropologist, explorer, and statistician in late 1800 [17]. A questionnaire is used in case where resources are limited, as a questionnaire can be quite inexpensive to design and administer. In this case, time is an important resource which a questionnaire consumes to its maximum extent. Also, protection of privacy of the participants is essential, as participants will respond honestly only if their identity is hidden and confidentiality is maintained. Moreover, questionnaires can be useful confirmation tools when corroborated with other studies that have resources to pursue other data collection strategies [18]. In this paper, data were collected using a questionnaire from 115 different laboratories in Baghdad, Iraq, to assess the specific incidence rank of 15 infectious diseases and diabetes in the Iraqi community during 2019, followed by another questionnaire to identify how to diagnose these diseases. Consequently, applying various machine learning techniques for the diagnosis of the highest spread diseases in the Iraqi community to automate this process to evaluate the performance of this framework. However, all the techniques applied performed well, and compared to various measures, they achieved excellent accuracy.

2. Proposed method

The proposed questionnaire framework was implemented in two parts, the first part includes questionnaires that were conducted; first, the most common diseases questionnaire (MCDQ) followed by symptoms questionnaire (SQ). So far, questionnaires data were collected for MCDQ a 15 infectious diseases and diabetes in Iraq to rate the highest spreading diseases in the community in 2019. However, popularity was rated scale by most laboratories in Baghdad province. Then, each laboratory wrote its own rate for each disease. The case definitions of the diseases were determined according to the International Classification of Diseases Tenth Revision (ICD-10) usually using tools depending upon clinical examination and microbiological and/or radiological tests as needed. After that, the case was classified either as definite, probable, or possible, and mostly, only definite cases were reported. SQ identified the most symptoms related to each disease by mean all symptoms that can be observed on the patent, although, blood tests have been recognized. So far, another part was conducted to evaluate this questionnaire framework efficiently to gain an overall perspective to a proposed method for medical diagnosing especially infectious disease diagnoses using various machine learning algorithms. This paper has highlighted the questionnaire framework through an explanation of each stage of questionnaires passed. In order to show how fit this framework is to support healthcare specialists by employing techniques of machine learning, Figure 1.
2.1. Questionnaires design
This step is crucial to ensure the time spent on designing, managing, and tabulating questionnaire data contributing to reliable and usable results. Educators should follow research-based principles for creating questionnaires [19]. A well-designed questionnaire requires thought and effort and needs to be planned and developed in several stages as illustrated in Figure 2 (Steps of design a questionnaire). There are four different types of questionnaire designing, which could be applied to these types according to the purpose of the questionnaire (Contingency questions/Cascade format, Matrix questions, Closed-ended questions, and Open-ended questions). Closed-ended questions are popular because the data generated can be summarized and analyzed more efficiently than the data from open-ended questions. Closed-ended questions are often used when time and resources for analysis are limited [17]. Also provide limited response options that contain different types (Single response, Checklist, Ranking, Rating scale) [20]. Therefore, a closed-ended question has been applied, as a response method rating scale type used for collecting responses from specialists.
2.2 Methods of collecting target response
There are four main types of collecting response as the following:
1. Face-to-face interview.
2. Telephonic interview.
3. Mail questions.
4. Internet questions.
In this questionnaire, the first approach was used, which is personal interviews. It is a survey performed in person by an interviewer that normally goes to the person being surveyed. This approach can have high response rates and can answer questions if appropriate, to influence the collection of respondents to use longer, more nuanced questionnaires to encourage respondents. The disadvantages of this approach may also be in the form of high costs, time-consuming, more variety of administrative criteria and training interviewers, traveling, and contacting respondents [18].

2.3 MCDQ
This questionnaire is essentially needed because there are limited resources for specifying the most common infectious diseases in the Iraqi community. Therefore, a questionnaire should be designed carefully. In the first step, a questionnaire developer should determine the aim of the questionnaire [21]. In this study, the main aim is to specify the most popular diseases in the Iraqi society and understanding the nature of spreading diseases in that specific demographic area. Consequently, on the questionnaire output, the scope of the study can be clear and determined. In the second step, needed information is initiated by understanding each goal of the questionnaire as explained in the first step. Therefore, the questionnaire consists of 15 infectious diseases named (Influenza (Flu), Typhoid, Hepatitis, Tuberculosis, German measles, Cholera, Meningitis, Allergic conjunctivitis, Smallpox, Scabies, Chickenpox, Elephantiasis, Malaria, Rabies, SARS) as well as diabetes as shown in appendix Questionnaire form. Each row represents diseases while each column describes the degree values depending on spreading in society from 0 to 5. In the third step, each of these diseases has a rating scale of 1 to 5. After that, we asked 115 laboratories to rate 1-5 for each disease, based on the disease popularity in the Iraqi community. In the fourth step, the draft questionnaire had been reviewed by people who are not involved in its creation to ensure that the questions align with the goals and information needed as mentioned in the first step. In the final step the content of the questionnaire was ready and the focus shifted to organizing and formatting it to match the goals of the questionnaire. MCDQ was filled by specialists by asking them to rate each disease from 1-5. Then descriptive
statistical analyses were carried out. This analysis is represented by calculating the average of complete responses for each one of the pre-dedicated diseases. For example, diabetes disease weight was determined by summing up all the response values separated by the number of samples obtained (115). Accordingly, the weight of each disease can be calculated based on the subsequent equation:

\[
\text{Disease\_Rank} = \frac{\sum \text{samples\_responses}}{N}, \text{ where } N \text{ represents some samples}
\] (1)

Equation 1 was used to examine the differences between proportions value of 1-5. Figure 3 below shows how the mean of each disease is valued. The primary analyses were stratified by demographic characteristics and degree of diseases spreading, depending upon 115 questionnaire samples collected from different demographic areas in Iraq for example Baghdad, from different places for the laboratories such as Adhamiya, Al-Mansur, Karada, etc, showing that the top three diseases were diabetes (4.79), flu (4.71), and typhoid (4.31). These values represent the mean of spreading degree.

![Diseases Rank](image)

**Figure 3.** Diseases rank.

### 2.4 SQ

Moreover, it can be shown that there is minimal work on infectious diseases in Iraq and that there is no research on flu and typhoid among the Iraqi population. For this reason, a second questionnaire was conducted to specify how to diagnose the highest spread of diseases (diabetes, flu, and typhoid) including symptoms and blood tests. All these features can help to diagnose these diseases. The steps below show how to design SQ:

- In the first step, the main purpose of this questionnaire was to define and estimate the symptoms and diagnostic process of each of the three diseases (diabetes, flu, and typhoid).
- The second step was preparing information required for the questionnaire, consisting of three questions.
  
  In order, each question asking for describing the diagnosis of disease by symptoms and/or blood tests. By checking symptoms then should mention all symptoms observed for the disease further any additional information, however, by checking blood test should mention the normal range of each test required also if there any additional information must be written.
- In the third step sequence of questions depending on the output of questionnaire one which is diabetes, flu, and typhoid.
- In the fourth step, a draft copy of the questionnaire was reviewed by a specialist to keep alignment with the goal in the first step.
- In the final step, the questionnaire content was reviewed and ready for applying formats to finalize the questionnaire.

However, laboratory specialists were asked to indicate symptoms of each disease that have been identified as well as blood test factors with normal ranges of these factors. The total number of collected responses was 115 samples. Only a small number of respondents indicated insufficient
information for analysis steps of diagnosis. Table (1) describes disease symptoms for each disease and illustrates blood test factors of diagnosis diseases.

Table 1. Diseases symptoms and blood tests.

| Symptoms     | Flu       | Typhoid  |
|--------------|-----------|----------|
| Diabetes     | Fever     | Fever    |
| Polyuria     | Headache  | Headache |
| Sweating     | Runny nose| Abdominal pain |
| Losing weight| Headache  | Diarrhea |
| Thirst       | Sneezing  | Cough    |
| Fatigue      | Cough     | Cough    |
| Stress       | Tiredness | Sweating |
| Nausea       | Fatigue   | Fatigue  |
|              |           | Vomiting |

| Blood Test   |           |          |
|--------------|-----------|----------|
| HB1AC        | ESR       | Widal test |
| FBS          | WBC       |           |
| Glucose urine| Glucose tolerance test | |

Table 1 shows the symptoms for each disease and blood tests factors. Diabetes has seven symptoms and so does flu, whereas typhoid has eight symptoms. Also, Table 1 demonstrates blood test factors which were considered as features used to support the diagnose process of these diseases. Firstly, HB1AC means a form of hemoglobin (a blood pigment that carries oxygen) that is bound to glucose. The blood test for the HbA1c level is routinely (3 months) performed in people with type 1 and type 2 diabetes mellitus. Its normal range is <4%-6%. RBS means a random blood sugar test which examines your blood glucose at a random time of the day (after eating 2 hours). A normal range between 100-140 mg/dL (milligrams per deciliter) or higher is a sign that you have diabetes. On the other hand, FBS means Fast Blood Sugar test to determine how much glucose (sugar) is in a blood sample after an overnight fast (at least 8 hours). Its normal range is <75-120 mg/dL>. Glucose urine measures the amount of sugar (glucose) in a urine sample. It can be only positive or negative. Additionally, the Glucose tolerance test measures how well your body cells can absorb glucose (sugar) after you consume a specific amount of sugar. So far, flu disease blood test factor includes ESR which means erythrocyte sedimentation that measures and tracks inflammation in the body. It measures how easily erythrocytes (red blood cells) settle at the bottom of the blood sample test tube if the size is more than 15 ml/1hr, suggesting that there is an inflammation. On the other hand, WBC means White Blood Cell, which indicates mostly within a natural variety of (4000-11000). Even though the Widal test is a probable serological test for enteric fever whereby bacteria that cause typhoid fever are combined with a serum containing particular antibodies from infected individuals. It can be either positive or negative.

2.5 Dataset describing
Both datasets of flu and typhoid contain 100 responses (sample) collected through a tailored questionnaire. However, the datasets features are referred to earlier in Table 1. Datasets were used as input data to diagnose patients, based on whether or not they are infected. Table 2 displays the characteristics of the datasets used in this study.

Table 2. Datasets characteristics.

| Data Set | No. Samples | No. Features | No. Classes |
|----------|-------------|--------------|-------------|
| Flu      | 100         | 9            | 2           |
| Typhoid  | 100         | 9            | 2           |

Flu disease has been shown to typically diagnose seven symptoms and two blood test features, while typhoid has eight symptoms and one blood test feature. These symptoms are typically widespread in
Iraqi environment, And it can be verified by the WHO [22,23], NHS [24,25], CDC [26,27], and ECDC [28,29].

3. Experimental results and analysis
This paper uses the MATLAB R2020a platform to diagnose diseases (flu and typhoid) as an example of assessing the questionnaire framework proposed. The experimental results show a good framework questionnaire for models of machine learning by achieving high accuracy of classification through different techniques of machine learning each had its strategy to classify the samples. These implemented models were evaluated by using five K-fold cross-validations to improve the strength of these machine learning classifiers. Five K-fold cross-validations were used for dataset training to avoid over-fitting, which implies that the classifier will not generalize well from our training data to various datasets. The dataset acquired is first split into 5 equalized subsets. That one was carried out as a test dataset, while the classification technique was trained on all remaining samples and an identical number of samples was randomly selected from four datasets. In this paper, machine learning techniques were used to test the proposed questionnaire framework. These ML techniques of Naive Bayes, Linear discriminant Ensemble (subspace discriminant), and Quadratic discriminant techniques were used for flu disease. While KNN (weighted KNN) and native Bayes (kernel native Bayes) were used for typhoid disease. Tables 3 and 4 show and describe the performance of these classifiers. Consequently, the effectiveness of the classifiers models were accurately measured for performance by computing the numbers of true and false samples that have been classified. Accuracy of the machine is calculated by the following formula:

\[
\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}
\]

Where; TP is true positive which means we predicted yes (they have the disease), and they truly have the disease, TN is true negative, which means that we expected no, and the patients were not affected, FP means false positive, which means that we expected yes, but they do not have a disease, and FN means false negative, which means that we expected no, but they do have a disease.

| Model                          | Accuracy | Training time (Sec) | Misclassification samples |
|-------------------------------|----------|---------------------|--------------------------|
| Native bayes                  | 92.0%    | 2.7477sec           | 8                        |
| Linear discriminant           | 92.0%    | 1.9866sec           | 8                        |
| Ensemble (subspace discriminant) | 92.0%   | 8.1362sec           | 8                        |
| Quadratic discriminant        | 91.0%    | 1.8343sec           | 9                        |

Table 3. Machine learning techniques performance for flu disease.

| Model | Accuracy | Training time (Sec) | Misclassification samples |
|-------|----------|---------------------|--------------------------|
| KNN   | 98.0%    | 17.453sec           | 2                        |
| Native bayes | 84.0%    | 13.362sec           | 16                       |

Table 4. Machine learning techniques performance for typhoid disease.

As seen in the tables above, the precision, training time, and misclassifying samples were measured where there are three models had the highest performance established for the classification method of the flu disease classification models was 92 % which is different from each other in the mechanism of each algorithm and training time. On the other hand, the high-grade performance of typhoid fever is 98 % which is KNN. Figures (4,5) defines the accuracy of the flu and typhoid disease classifier models, which contributes to the high accuracy of the classifiers, the strength, and the effectiveness of the proposed structure.

As aforementioned before that native Bayes is beside the other two models in table 4 had the same accuracy, we selected it as an example model native Bayes because it performs well to diagnose flu disease while performing to diagnosing typhoid disease not the same because the KNN model performs excellently to diagnose patients because of achieving the highest performance accuracy as shown in figure 5, 6 and 7 show Receiver Operating Characteristic (ROC) which measures the output
of the classifier over a sequence by the trade-offs between the sensitivity and the specificity of each cut-off. Although the area under the curve (AUC) has been generated, it depends on the expected outcome. Furthermore, they show the confusion matrix for native Bayes and KNN model. For native Bayes, CM indicates that there is 48 samples predicted with no flu disease while 44 were predicted to have the flu. Also, 8 misdiagnosed samples occurred, 5 of them had the flu and were diagnosed as not infected, and 3 of them had no flu and were diagnosed as infected ones. On the other hand, KNN CM indicated that there are 52 samples diagnosed with no typhoid infect and 46 samples were predicted as infected. In addition, only 2 were misdiagnosed as no typhoid infect while they in fact were.

Figure 4. Accuracy of flu disease using machine learning techniques.

Figure 5. Accuracy of typhoid disease using machine learning techniques.

Figure 6. ROC and Confusion matrix of Native bayes for flu disease.
Figure 7. ROC and confusion matrix of KNN for typhoid disease.

4. Conclusion
Infectious diseases have a significant impact on the behavior of Iraqi society and population. This leads to a big problem for the healthcare specialists on how to diagnose these diseases correctly or mortality could be the consequence. Automated accurate diagnosis helps saving time, effort, and cost for the patients. In this paper, the questionnaire framework has been expressed in details by designing a it for acquiring information which shows that the highest value assigned to diabetes was not an infectious disease but it spreads rapidly between the populations in a rate of 4.79. The second disease is the flu which is a seasonal influenza type of cold with a value of 4.71. The third was assigned to typhoid fever whose value was 4.31. Then it is followed by identified overall features related and used later as input for classifying diseases using ML techniques. The questionnaire framework produced is considered a starting point for the researchers in the field of medical diagnostics utilizing artificial intelligence, starting with gathering information on the disease and the identification of the necessary features, and ending with the use of machine learning techniques. From statistics, it is shown that the three models (native Bayes, Linear discriminant, and Ensemble (subspace discriminant)) provide the best model accuracy for diagnosing flu disease, while the KNN model shows the best model with a significant accuracy of 98% for diagnosing typhoid disease. This work can be extended and improved to include the other highest widespread infectious diseases in the Iraqi community and include some other ML techniques for medical diagnosis.

5. References
[1] Nnadi C, Etsano A, Uba B, Ohuabunwo C, Melton M, Wa Nganda G, Esapa L, Bolu O, Mahoney F, Vertefeuille J, Wiesen E and Durry E 2017 Approaches to Vaccination Among Populations in Areas of Conflict (J. Infect. Dis) vol 216 pp 68–72
[2] World Health Organization 2018 The Top 10 Causes of Death (https://www.who.int/en/news-room/fact-sheets/detail/the-top-10-causes-of-death)
[3] Bonita R, Beaglehole R and Kjellström T 2006 Basic Epidemiology Gynaecol Oncol. MRCOG Beyond, Second Ed.
[4] Gould D W, Walker D and Yoon P W 2017 The Evolution of BioSense: Lessons Learned and Future Directions (Public Health Rep.) vol 132 no 1 pp 7S-11S
[5] A Signore 2013 About Inflammation and Infection (EJNMMI Res.) vol 3 no 1
[6] Health N I 2017 Understanding Emerging and Re-Emerging Infectious Diseases (NIH Curriculum Supplement Series)
[7] Miceli A, Sebuyira L M, Crozier I, Cooke M, Naikoba S, Omwangangye A P, Rayko-Farrar L, Ronald A, Tumwebaze M, Willis K S and Weaver M R 2012 Advances in clinical education: A
model for infectious disease training for mid-level practitioners in Uganda (International Society for Infectious Diseases) vol 16 no 10
[8] Kenyon Lischer S 2008 Security and Displacement in Iraq: Responding to the Forced Migration Crisis (International Security) vol 33 no 2 pp 95–119
[9] Eiset A H and Wejece C 2017 Review of Infectious Diseases in Refugees and Asylum Seekers-Current Status and Going Forward (Public Health Rev.) vol 38 no 22 pp 1-16
[10] C A Mattmann 2013 Computing: A Vision for Data Science (Nature) vol 493 no 1 pp 473-475
[11] V Marx 2013 The Big Challenges of Big Data (Nature) vol 498, no 1 pp 255-260
[12] I Kononenko 2001 Machine Learning for Medical Diagnosis: History, State of the Art and Perspective (Artif. Intell. Med.) vol 23 no 1 pp 89-109
[13] M Pradhan 2014 Data Mining& Health Care: Techniques of Application (Int. J. Innov. Res. Comput. Commun. Eng.) vol 2 no 12 pp 7445–7455
[14] T M Mitchell 1997 Machine Learning (McGraw-Hill Science/Engineering/Math) vol 17
[15] S R Salkuti 2020 A Survey of Big Data and Machine Learning (International Journal of Electrical and Computer Engineering) vol 10 no 1 pp 575–580
[16] A Selwal and I Raoof 2019 A Multi-Layer Perceptron Based Intelligent Thyroid Disease Prediction System (Indonesian Journal of Electrical Engineering and Computer Science) vol 17 no 1 pp 524–533
[17] C Chou, Y F Chang and Y Y Jiang 2000 The Development of an Online Adaptive Questionnaire for Health Education in Taiwan (Computers and Education) vol 35 no 3 pp 209-222
[18] Shigeru Omi, H Ealth, R Esearch and M Ethology 2001 A Guide for Training in Research Methods (World Health Organization)
[19] T F Burgess 2001 A General Introduction to the Design of Questionnaires for Survey Research (University of Leeds) vol 20 no 3
[20] W Yuan and Y Zeng 2017 Study of Methods to Improve the Counselors’ Scientific Research (Sci. Res. Publ.) vol 8
[21] E Fanning 2005 Formatting a Paper-based Survey Questionnaire: Best Practices (Pract. Assess. Res. Eval. peer-reviewed Electron. J.) vol 10 no 12
[22] World Health Organization 2018 Typhoid (https://www.who.int/news-room/factsheets/detail/typhoid)
[23] World Health Organization 2018 Influenza (Seasonal) (https://www.who.int/news-room/factsheets/detail/influenza-(seasonal))
[24] National Health Service 2019 FLU (https://www.nhs.uk/conditions/flu/)
[25] National Health Service 2018 Typhoid Fever (https://www.nhs.uk/conditions/typhoid-fever/symptoms/)
[26] Centers for Disease Control and Prevention(CDC) 2019 Influenza (Flu) (https://www.cdc.gov/flu/symptoms/index.html)
[27] Centers for Disease Control and Prevention (CDC) 2019 Typhoid Fever and Paratyphoid Fever (https://www.cdc.gov/typhoid-fever/symptoms.html)
[28] European Centre for Disease Prevention and Control and (ECDC) 2019 Factsheet about seasonal influenza (https://www.ecdc.europa.eu/en/seasonal-influenza/facts/factsheet)
[29] European Centre for Disease Prevention and Control and (ECDC) 2019 Facts about typhoid and paratyphoid fever (https://www.ecdc.europa.eu/en/typhoid-and-paratyphoid-fever/facts)