ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

An Interactive and Predictive Pre-diagnostic Model for Healthcare based on Data Provenance

Zhwan Namiq Ahmed[1], Jamal Ali Hussien[2]

[1]Department of Computer Science, College of Science and Technology, University of Human Development, Sulaymaniyah, Iraq, [2]Department of Computer Science, College of Science, University of Sulaimani, Sulaymaniyah, Iraq

ABSTRACT

The future of health care may look completely different from the current clinic-center services. Rapidly growing and developing technologies are expected to change clinics throughout the world. However, the health-care delivered to impaired patients, such as elderly and disabled people, possibly still requires hands-on human expertise. The aim of this study is to propose a predictive model that pre-diagnose illnesses by analyzing symptoms that are interactively taken from patients through several hand gestures during a period of time. This is particularly helpful in assisting clinicians and doctors to gain a better understanding and make more accurate decisions about future plans for their patients' situations. The hand gestures are detected, the time of the gesture is recorded, and then they are associated with their designated symptoms. This information is captured in the form of provenance graphs constructed based on the World Wide Web Consortium provenance data model. The provenance graph is analyzed by extracting several network metrics and then supervised machine learning algorithms are used to build a predictive model. The model is used to predict diseases from the symptoms with a maximum accuracy of 84.5%.

Index Terms: Hand Gesture Detection and Recognition, Pre-diagnosis Disease, Data Provenance, Provenance Network Analytics, Machine Learning Algorithm

1. INTRODUCTION

As technology has become a part of human's life for decades, the study of the relationship between humans and computing technology in so-called human-computer interaction (HCI) is essential to serve the human's needs [1]. While many developed countries gradually step into population aging, many researches on elderly people have been conducted, especially about the interaction with computers for seniors [2]. The number of disabled people is also rising due to wars [3]. Therefore, it is really imperative to consider how computing technology will be able to meet the needs of these important users. Moreover, with the widespread popularity of computing technology in modern society, information technology becomes continuously incorporated into the daily lives of people [2], [4]. Nowadays, the use of technology such as computers in health care is significantly developing and growing, becoming an essential part of clinical services [5].

With the development of ubiquitous computing technology, end-user interaction approaches with traditional mouse and keyboard and electronic pen are not sufficient [6]. Therefore, we have to think about other ways to interact or communicate with computing technologies, especially for these two types of users (elderly and disabled) who may not

Access this article online

DOI:10.21928/uhdjst.v3n2y2019.pp59-73 | E-ISSN: 2521-4217 P-ISSN: 2521-4209

Corresponding author's e-mail: Zhwan N. Ahmed, Department of Computer Science, University of Human Development, Sulaymaniyah, Iraq. E-mail: zhwan.ahmed@uhd.edu.iq

Received: 26-08-2019 Accepted: 24-09-2019 Published: 01-10-2019

be able to communicate with technologies normally due to several body problems that limit their movements. There are several gestures that can be used to interact with technologies, such as head, face, hand, and finger movements [4], [5]. These gestures might be used in various fields, namely health care [7], speaking, listening [8], gaming [9], and many others. In this research paper, hand gesture is chosen instead of sign language to make a communication between these two types of users and computing technologies since it has the advantages of simplicity, requiring less learning time, and it can be attractive and make applications more accessible [1]. However, sign language is more complex for elderly and disabled patients because it requires knowledge and more training to learn.

The World Wide Web Consortium (W3C) standardized provenance (PROV) data model defines data provenance as a technique that describes the origin and history of data [10]. Data provenance has been successfully employed in a variety of applications, including description of patient's historical information [11]. The primary role of data provenance is to connect pieces of data with the processes that have made the data and to document how data are transformed during their life cycle [12]. This paper used provenance graphs to represent relationships between symptoms and related diseases, and also to record patients' history information, because according to Asan and Montague [7] and Shickel *et al.* [13], recording health information provides a great potential that can be used to improve the quality of care. In addition, in health-care data setting, provenance is able to deliver audibility and transparency and accomplish trust in software systems [11]. Machine learning classification algorithms are currently well appropriate for analyzing medical data. Furthermore, there is a lot of work done in medical diagnosis for specific diagnostic problems [14].

In this paper, we propose a system that is capable of predicting pre-diagnosis of potential diseases based on data provenance with machine learning algorithms. The main idea is to help clinicians in Kurdistan Region of Iraq (KRI) to monitor and take care of elderly and disabled patients. In particular, this model is helpful for physicians to make better, faster, and more accurate decisions about patients' health conditions. The pre-diagnosis process is conducted using a series of symptoms provided by the user using a set of hand gestures. Each hand gesture is associated with a symptom. Different patterns of hand gestures are captured and processed by an HCI sub-module. The hand gestures, the time of their occurrence, and their relationships with the patient and the diseases are stored as provenance graphs. Network metrics are extracted from these provenance graphs as aggregate information. This information is fed to a prediction model that uses three supervised machine learning algorithms, namely, decision tree classifier (DTC), K-nearest neighbor (KNN), and support vector machine (SVM). The highest accuracy rate is produced by the DTC algorithm with approximately 84.5%, which is higher than the results of prior studies that are discussed in experimental results section.

The rest of the paper is organized as follows: The related work on designing medical applications based on data provenance with machine learning, and diagnosis diseases are reviewed in section 2. Section 3 provides a theoretical background which gives an explanation for the conceptual view of data provenance, and introduction for different types of machine learning algorithms. In Section 4, the study method for the proposed work is illustrated by providing the recognition of hand gestures, representing provenance graphs, and examining machine learning algorithms. Constructing the dataset of the system and building a predictive model are discussed in Section 5. Experimental results are discussed in Section 6. Finally, in Section 7, the conclusions and directions for future work are presented.

2. THEORETICAL BACKGROUND

In this section, provenance and several types of supervised machine learning algorithms are explained.

2.1. Conceptual View of Data Provenance

Data provenance is the technique that illustrates what influenced the generation of an object that would be physical, digital, or conceptual [11]. It has become a very significant topic in many scientific communities since it displays the data movement in systems [15], [16]. Furthermore, given information about the place data originated from, how they come in their present states, and who or what acted on them helps users to establish trust in the data. Provenance can show resources and relations that have affected the construction of the output data and are commonly expressed as directed graphs (digraphs) [17]. The primary aim of the W3C standardized provenance is to enable the extensive publication and exchange of provenance over the web [18]. The provenance data model is subject to a set of constraints and inference rules [19], which are useful in validating the provenance information [20] and are essential for preserving graphs integrity when converting provenance graphs [21], [22]. To establish trust of data, these properties must be maintained when capturing provenance information and constructing provenance graphs [23].

The graph-based model is built on three main concepts, namely, entity, activity, and agent. Entity can be defined as real or imaginary, or any kind of things, physical, or digital. Activity is a procedure that happens over a period of time and may affect the state of entities and generate new ones. The agent is something that takes some form of responsibility for an activity, or for another agent's process. The relationships that normally occur between entities, agents, and activities can be described by the means of the graph, as shown in Fig. 1.

Fig. 1 represents the core provenance data model components. Each activity can start and finish at a particular time; these processes are described using two relations: *prov: startedAtTime* and *prov: endedAtTime*. In addition, during activities lifetime, they can use and generate a variety of entities, which are presented with *prov: used* and *prov: wasGeneratedBy,* respectively. To provide some dependency information, an activity *prov: wasInformedBy* another activity without providing activities' start and end times. *prov: wasDerivedFrom* relation can be used to form a transformation from one entity to another. In addition, agents have responsibilities for any activity and entity within provenance which is described using relations: *prov: wasAssociatedWith* and *prov: wasAttributedTo,* respectively. Finally, agents can be responsible for other agents' actions which express as the influencing agent *prov: actedOnBehalfOf* another.

2.2. Machine Learning Algorithms

Machine learning algorithms are a group of algorithms or processes that help a model to adapt to the data [24]. Furthermore, machine learning algorithms usually specify the way the data are transferred from input to output, and how the model learns the appropriate mapping from input to output. The following are three types of supervised machine learning algorithms that are introduced.

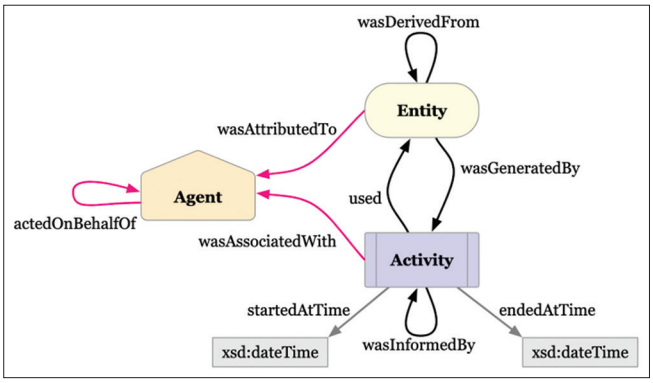

Fig. 1. Core provenance data provenance [16] (Relationships between entity, agent, and activity).

2.2.1. DTC

It is a supervised learning algorithm that is typically applied for classification problems. DTC breaks down data into small subsets at the same time, a related decision tree will be incrementally developed [24].

2.2.2. KNN

It is a simple algorithm that stores all accessible cases and groups new cases by a majority vote of its "K" neighbors [25], [26]. The case being allocated to the class is the most common among its KNN measured by distance functions, which would be "Manhattan," "Euclidean," "Minkowski," and "Hamming Distance" [25].

2.2.3. SVM

SVMs are well-known supervised classification algorithms that separate various groups of data [27], [28]. These vectors are grouped by optimizing a specific line so that the neighboring point in each group will be the farthest away from each other. SVM can be used for both regression and classification problems [29], [30].

3. RELATED WORK

Huynh *et al.* [12] proposed a number of network metrics for provenance information, and then they applied machine learning techniques over metrics to build predictive models for several key properties of the data. They adopted the provenance data model in their analytics. They discovered 22 metrics as features that were presented as a generic and principled data analytics method for analyzing data provenance graphs. Then, they used this method through the DTC technique to construct predictive models based on the provenance information. Moreover, they applied this method on the real-world data from three different applications, which, respectively, conducts the owner's identification of provenance documents, evaluates the quality of crowdsourced data, and describes the instructions from chat messages in an alternate reality game.

Choudhury and Gupta [14] presented diabetes disease detection through applying five supervised machine learning techniques. The dataset contains a total number of 768 samples and 9 attributes. There are two types of classes in their study, which are: Positive class (shows the person is diabetic) and negative class (indicates the person is not diabetic). In addition, their study presents a comprehensive comparative study on several machine learning algorithms based on a number of parameters, such as accuracy, recall,

precision, and specificity. Eventually, they found that the Logistic Regression algorithm provides the best accuracy result to classify the diabetic and non-diabetic samples.

Enriko [31] performed a comparative study of heart disease diagnosis systems using ten data mining classification algorithms, such as DTC, SVM, KNN, Naive Bayes, and Logistic Regression. Medical record data are gathered from a cardiovascular hospital. Based on several parameters that were taken from patients, including blood pressure, chest pain, shortness-of-breath, palpitation, and cold sweat, machine learning algorithms are used to analyze a sample of cardiovascular patients' data and predict the heart disease type they may suffer. After applying the algorithms, they stated that the KNN algorithm provides attractive results since it is always one of the top three algorithms for accuracy and performance.

The diagnosis of liver disease at the primary stage is essential for better treatment. However, it will be very difficult task for doctors if the symptoms become apparent when it is too late. To overcome this issue, Lavanya *et al.* [32] proposed a method that can be used for diagnosing liver diseases in patients using machine learning algorithms. The techniques were used include SVM, Ada Boost, DTC, and Naive Bayes. Furthermore, the major objective of their work is to use different classification algorithms to identify patients who have liver disease or not. They used a multivariate dataset that contains ten variables, namely, age, gender, total Bilirubin, and direct Bilirubin which are information about participants. All values are real integers. It contains records of liver patients up to 416 and non-liver patient records of up to 167. According to their results, the Ada Boost technique resulted highest accuracy of 75.6%.

Diabetes is a set of diseases in which the human body cannot control the level of sugar in the blood. Alaoui *et al.* [33] proposed a system to classify female patients into two groups: Having diabetes or not. They used a dataset that is related to diabetes in women of 21 years or older. The major aim of this paper is to make a comparative study between four machine learning classification algorithms, namely, Naive Bayes, Neural Network, SVM, and DTC, to select the best algorithm that classifies patients more accurately. They also used two different types of data mining tools: Weka and Oranges to run the selected algorithms [34]. They have made a comparison between selected classification algorithms. According to the accuracy results that they gained, SVM has the highest accuracy result, so they use it to classify the above two groups.

The above prior studies are the most related to the proposed system and also the results of some of them are compared with the results of the proposed system in section experimental result.

4. THE PROPOSED SYSTEM

In this section, our proposed method for pre-diagnosing diseases is discussed in detail. Patients can interact with the system using non-verbal communication to understand the patient's situation through five hand gestures. Each gesture represents a specific type of symptom, as shown in Fig. 2.

- One-finger hand gesture: It shows that the patient has a headache symptom.
- Two-finger hand gesture: This hand sign indicates that the patient has a shortness of breath problem.
- Three-finger hand gesture: It shows that the patient has a chest pain symptom.
- Four-finger hand gesture: Four-Finger hand sign represents that the patient has a vomiting issue.
- Five-finger hand gesture: Abdominal symptom represented through the five-finger gesture.

The steps of the proposed system are summarized in Fig. 3. The system performs various processes to identify the disease that the patient suffers from. The steps are as follows:

1. Reading, detecting, and recognizing hand gestures.
2. Relating each gesture to the symptom that it represents.

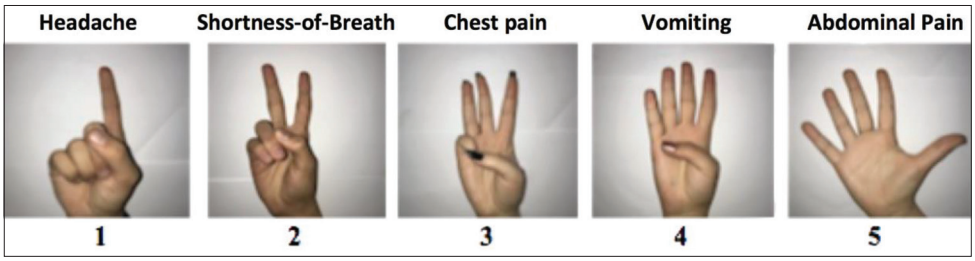

Fig. 2. Sign language digits dataset from 1 to 5 hand signs.

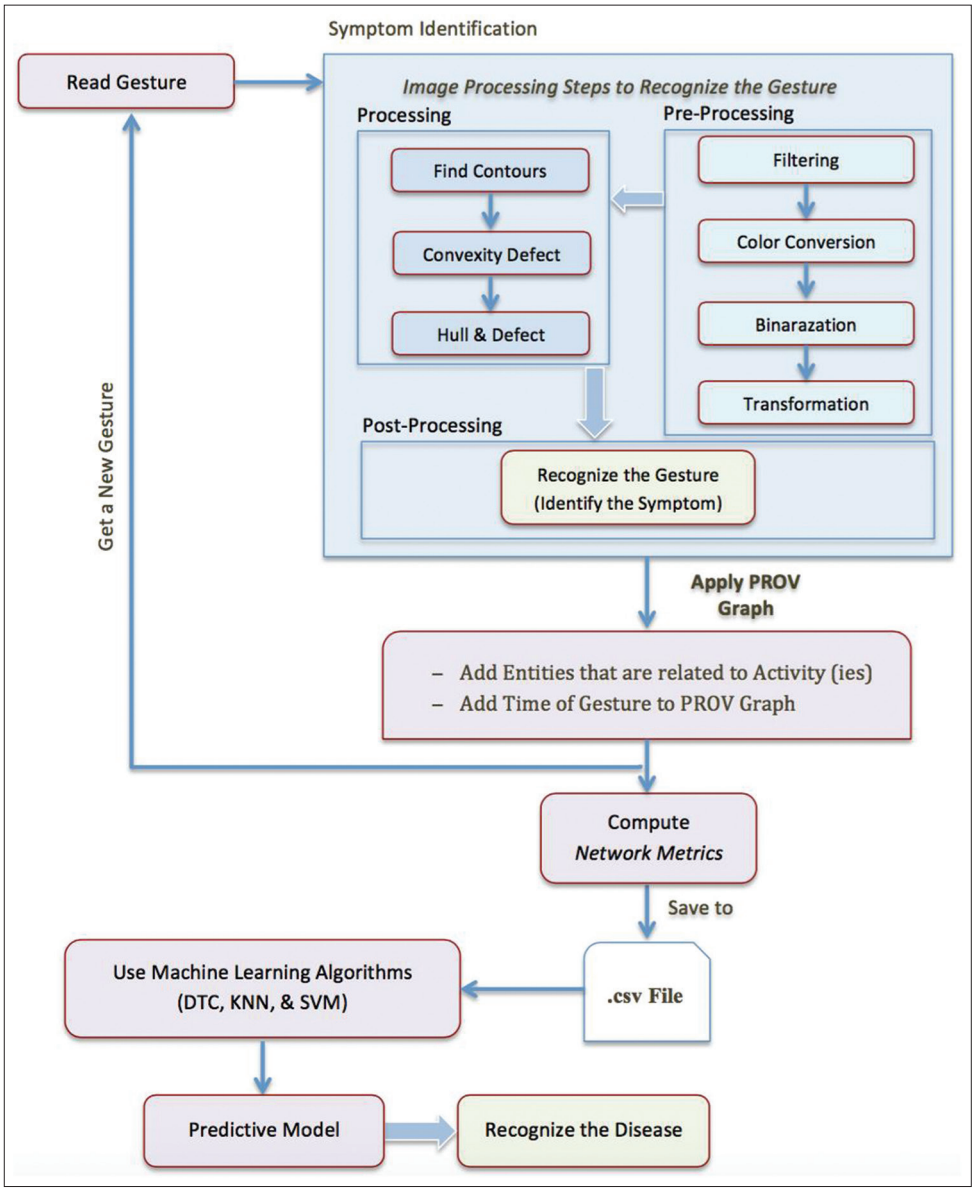

Fig. 3. Block diagram of the proposed system.

3. Constructing a provenance graph that represents the relationships among patients, symptoms, and diseases.
4. Extracting various network metrics from the provenance graph.
5. Using a variety of machine learning techniques to build a predictive model based on the network metrics.

4.1. Detecting and Recognizing Hand Gesture

The hand gesture recognition process is divided into three major steps, namely, pre-processing (Detection), processing (Feature Extraction), and post-processing (Recognition) [1], [35]. Each step works on the result of the previous step.

In the pre-processing detection step, several image filters, such as Gaussian blur, erosion, and threshold are implemented to detect the hand region of the images by converting the skin color of the hand to white and the rest of the image to black. Fig. 4 illustrates the results of the pre-processing step.

Feature extraction is implemented by the convexity defect technique on the images, which is initiated through discovering the contour of the hand. Then, the convex hull process is applied to describe the shape of the hand as a polygon, as shown in Fig. 5. After that, we find the variance between the contour and convex hull which defines as

convexity defect to find the number of defects of the hand gestures. As a result, these features are stored to be used in the last process to recognize hand gestures.

In the post-processing step, the numbers of defects that can be used to find the angle between two fingers are found. Furthermore, the parameter counter of the angle is calculated when the value of the angle is equal to or smaller than 90°, and the counter value is incremented until it reaches the maximum number of defects for one hand, which is four. The counter should be incremented by one for each gesture, since when the counter equals to zero, it means that the gesture has one finger, and it is true for all the other gestures.

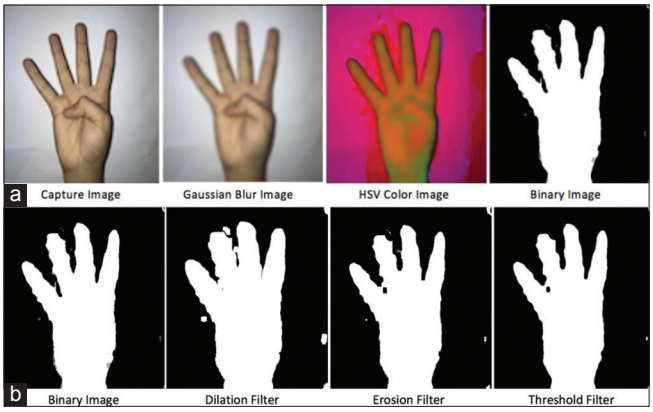

Fig. 4. The results of the pre-processing step. (a) Captured image from original to binary image, (b) applying several types of filtering on the binary image.

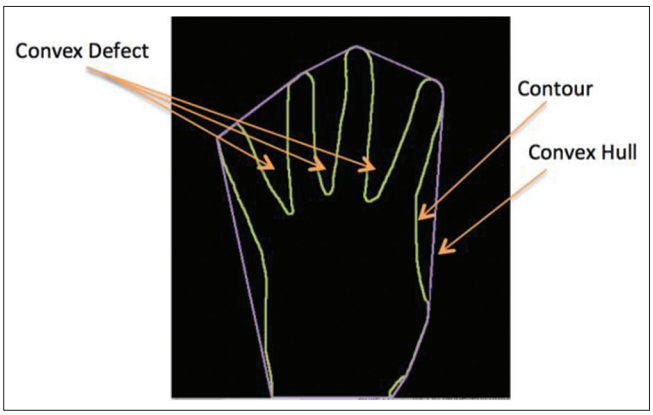

Fig. 5. Handshape with contour and convex defect.

4.2. Constructing Provenance Graphs Based on Different Types of Symptoms

4.2.1. Identification of symptoms and related diseases

A symptom, or side effect, is any illness, or mental change that is caused by a particular disease [36]. Health care experts use symptoms as clues that could help to analyze diseases. In this research work, we focus on five different types of symptoms: Headache, shortness of breath, chest pain, vomiting, and abdominal pain. Each symptom may relate to several diseases, as illustrated in Table 1 (all these data have been gathered by face-to-face interviews with three health care experts at different hospitals in KRI). A disease can cause more than one symptom, for example, anemia can cause headache and shortness of breath. Furthermore, two or more diseases may share the same symptoms, for instance, each of migraine, hypertension, and anemia can cause headache.

4.2.2. Constructing Provenance Graphs

The third step is the process of constructing a provenance graph from the symptoms. The system uses different provenance graphs for various symptoms; therefore, each symptom has a specific provenance graph that is different from the others. The provenance graph shown in Fig. 6 presents the headache symptom that is represented by one-finger gesture. When these symptoms occur, their provenance graphs will be merged into one single graph, which will be used for computing the network metrics.

On the one hand, it is possible for a user to enter the same hand gesture, which represents a specific symptom, many times repeatedly. On the other hand, there is possibility to have two or more different symptoms over a specific period of time. In the constructed provenance graph, symptoms are represented as activity nodes connected to entity nodes that represent the diseases. The patient who records the symptoms is represented as an agent node. The provenance graph shown in Fig. 7 illustrates a situation when a patient (the agent node *Patient: Ali*) suffered from symptoms represented

TABLE 1: Relationships between symptoms and diseases

Symptom	Disease		
Headache	Migraine	Hypertension	Anemia
Shortness of breath	Asthma	Myocardial infarction (MI)	Anemia
Chest pain	Pneumonia	Asthma	Myocardial infarction (MI)
Vomiting	Gastric ulcer	Migraine	Gastric
Abdominal pain	Gastric ulcer	Rental Stone	

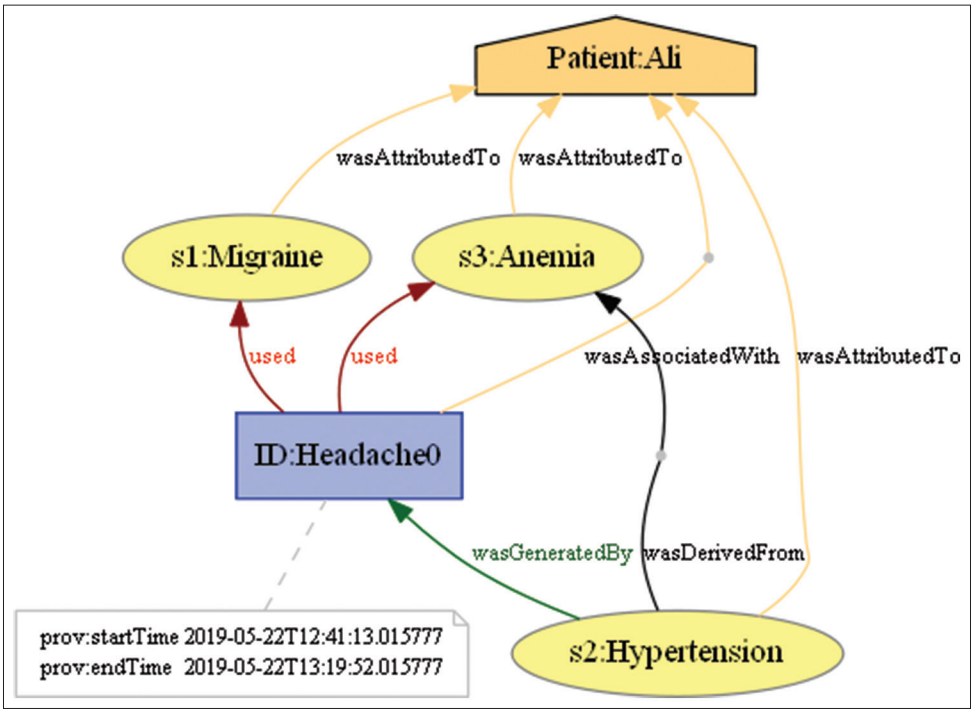

Fig. 6. Provenance graph for headache symptoms with its diseases.

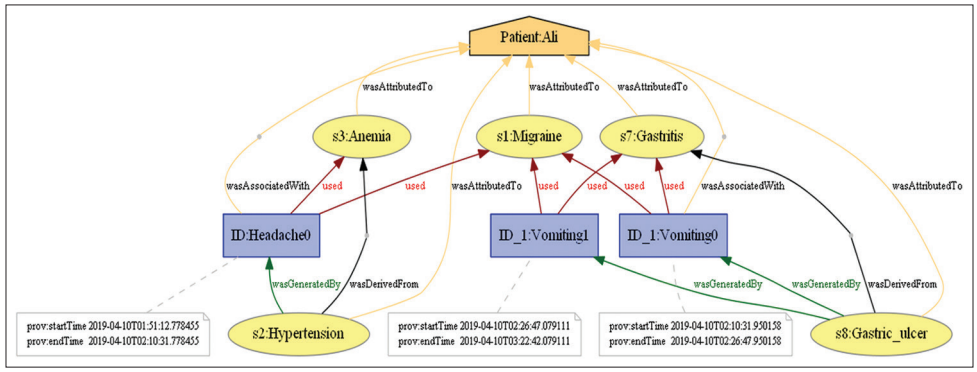

Fig. 7. Provenance graph for two types of symptoms with their diseases at specific duration time.

by the activity nodes *ID: Headache0*, *ID 1:Vomiting0,* and *ID 1:Vomiting1*. The start and end time of each activity node represents the period of time that each symptom lasts. Each symptom may affect or be affected by some diseases, which are depicted as the entity nodes *s1:Migraine*, *s2:Hypertension*, *s3:Anemia*, *s7:Gastritis*, and *s8:Gastric Ulcer*.

4.3. Provenance Network Metrics

Provenance network analytics is a novel data analytics approach, which can assist inferring properties of data, such as importance or quality, from their provenance Huynh *et al.* [12] and Roper *et al.* [15]. In the proposed system, several metrics are calculated by applying the equations that are presented below. These metrics are helpful in analyzing the patients' data and identifying patient disease.

The following equations are the metrics that are used in this research work:

- Time-duration (*TD*): It is the amount of elapsed time between the start and end of each similar symptom.

$$TD = EndStateTime - StartStateTime \quad (1)$$

In Fig. 7, there are two different types of symptoms: Headache and Vomiting. Using the start and end time of each symptom, the time durations (in minutes) of these two symptoms are:

$TD_{Headache}$ = (2:10:31) - (1:51:12) = 0:19:19 = 19 min

$TD_{Vomiting}$ = (3:22:42) - (2:10:31) = 1:12:11 = 72 min.

These indicate that the headache and vomiting symptoms lasted for 19 and 72 min, respectively.

- Similar symptom count (*SSC*): It is a number of similar symptoms (Hand Gestures) over a specific period of time.

So, $SSC_{Headache} = 1$, and $SSC_{Vomiting} = 2$ from 01:51:12 to 3:22:42.

- Symptom average time (*SAT*): It depends on the duration time and the number of similar symptoms. It represents the average time of the same gesture over a period of time and can be found using the following equation:

$$SAT = TD/SSC \quad (2)$$

Therefore,

$SAT_{Headache}$ = 19/1 = 19 min
$SAT_{Vomiting}$ = 72/2 = 36 min.

If a symptom does not exist in the provenance graph, then its $SAT = 0$.

- Total average time (*TAT*): It is the total of the average times of all the symptoms and it can be found using the equation below:

$$TAT = \sum_{k=1}^{5} SAT_k \quad (3)$$

Where each symptom is represented by a number k from 1 to 5, and SAT_k is the average time of the k^{th} symptom. In our example $TAT = 19 + 36 = 55$ min.

- Symptom percentage (*SP*): It is a ratio of the average time of a specific symptom (*SAT*) to the *TAT*. The system uses the *TAT* to find the percentage of each symptom using the following equation:

$$SP = (SAT/TAT) * 100 \quad (4)$$

Therefore,

$SP_{Headache}$ = (19/55) * 100 = 34.5%
$SP_{Vomiting}$ = (36/55) * 100 = 65.5%.

We compare these ratios that have been extracted from the provenance graph for each disease to other ratios, refer to as disease ratio, that have been gathered from three experienced physicians of two local hospitals in KRI. Table 2 presents the value of disease ratio for each disease with its symptoms.

- Percentage diseases ratio (*PDR*): It is a percentage of each disease, which has been taken from the percentage of the symptoms (*SP*) and disease ratios. We calculate it by the following equation:

$$PDR = (SP * Disease\ ratio)/100 \quad (5)$$

If a symptom is not related to a disease, then the percentage is zero. According to disease ratios are shown in Table 2, we find the *PDR* of all diseases that are related to the symptoms in the provenance graph of Fig. 7:

Headache symptom:

$PDR_{Migraine}$ = (34.5 * 60)/100 = 20.7%
$PDR_{Hypertension}$ = (34.5 * 30)/100 = 10.35%
PDR_{Anemia} = (34.5 * 10)/100 = 3.45%

Vomiting symptom:

$PDR_{Gastric\ Ulcer}$ = (65.5 * 35)/100 = 22.925%
$PDR_{Migraine}$ = (65.5 * 15)/100 = 9.825% ⇒ (9.825 + 20.7 = 30.525%)
$PDR_{Gastric}$ = (65.5 * 50)/100 = 32.75%

These diagnosis results show that the patient probably has Gastric disease due to the highest ratio.

5. BUILDING DATASET OF THE PROPOSED SYSTEM AND PREDICTIVE PRE-DIAGNOSTIC MODEL

In this research, Sign-Language-Digits-Dataset[1] [37], [38], [39] that contains ten different types of gestures is used. From that dataset, we use only five types of gestures (Fig. 2) to build

1 Sign-Language-Digits-Dataset is prepared by group of students who studied at Turkey Ankara Ayranc Anadolu High Schools in 2002.

TABLE 2: Symptoms with the value of disease ratios for each disease

Symptom	Disease					
	Disease ratio (%)		Disease ratio (%)		Disease ratio (%)	
Headache	Migraine	(60)	Hypertension	(30)	Anemia	(10)
Shortness of breath	Asthma	(50)	Myocardial infraction (MI)	(30)	Anemia	(20)
Chest pain	Pneumonia	(40)	Asthma	(20)	Myocardial infraction (MI)	(40)
Vomiting	Gastric ulcer	(35)	Migraine	(15)	Gastric	(50)
Abdominal pain	Gastric ulcer	(60)	Rental stone	(40)		

our dataset of metrics that are used as input to the machine learning algorithms. After recognizing the gestures, the system computes several network metrics about the diseases based on the symptoms, as discussed in the previous section. The dataset contains exactly 1123 records; then they divided into 15 fields which includes information on the patients' IDs and symptoms, network metrics, and types of diseases which patients suffer during a period of time, as shown in Table 3. We divided our dataset into two.*csv* files: training and test. The training file contains exactly 914 records, and it is used to build a predictive model that recognizes diseases. The test file, on the other hand, includes 209 records that are used to test the model.

In the last step of the proposed system, we choose three different supervised machine learning techniques to train our model [24], [25], which are used for both classification and regression problems [40]. We apply them in terms of classification problem since our data are split based on different conditions, the results of these algorithms are better for the proposed system compare the others, and according to previous studies these algorithms are applied in terms of diagnosis disease.

5.1. DTC

In our system, DTC builds a classification model that can be used to diagnose diseases. Data are classified into feature variables, which are the network metrics, and target variables that contain different types of symptoms. To create the predictive model, the system applies DTC in this specialization where criterion technique is entropy, which controls how DTC decides to split the data, and further it actually affects how DTC draws its boundaries. Furthermore, the tree max depth is five on the training data. The reason why we choose five for the maximum tree depth is that almost leaf nodes are obtained at Level 5. Besides, if the maximum tree depth is smaller or greater than five, the accuracy decreased. Next, the test .csv file is applied to the model to find the accuracy of the system. The visualization of the DTC in our system is depicted in Fig. 8.

Fig. 8 Illustrates the entire structure of the DTC. All the nodes, except for the terminal (leaf) nodes, contain four parts:

1. Questions asked about the data based on the value of a feature; the answer to each question is either true or false.
2. Average weighted entropy represents the impurity of a node and it should be decreased as we move down the tree.

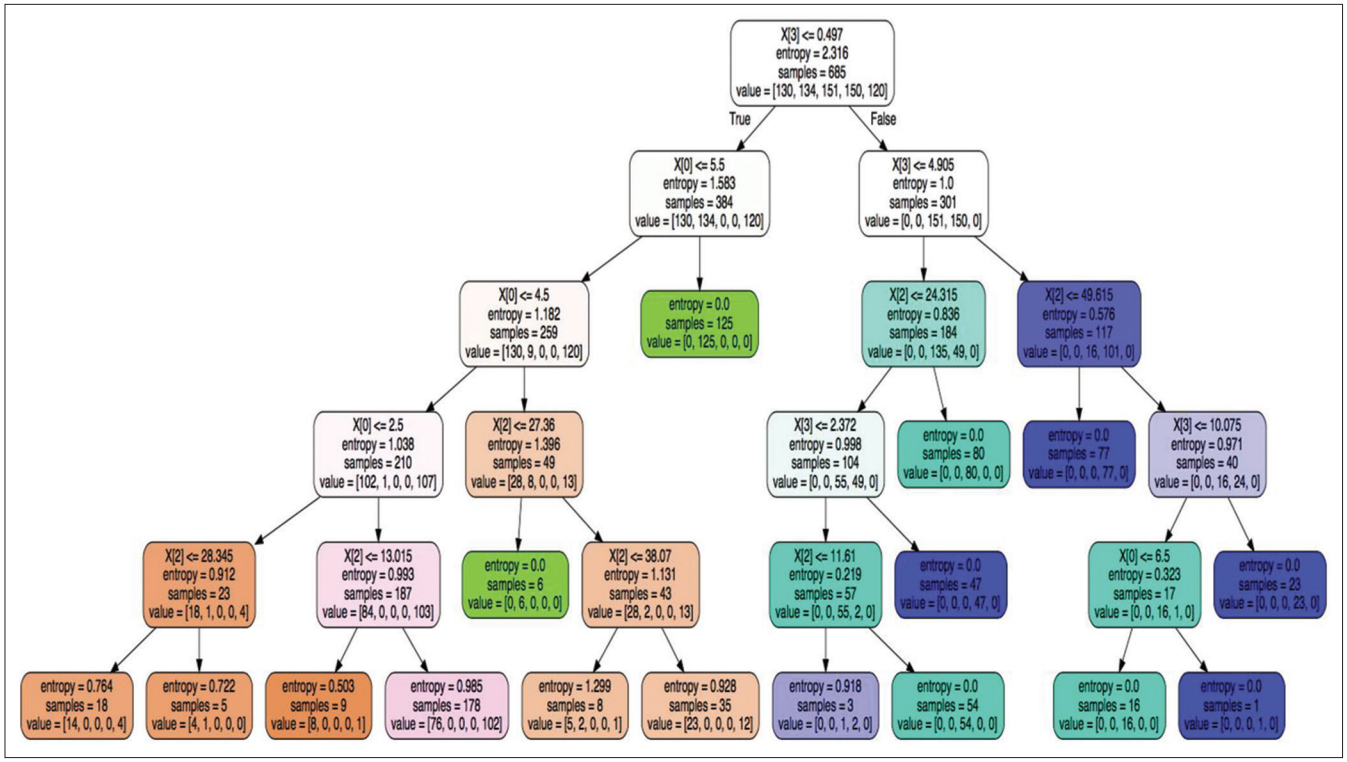

Fig. 8. Visualizing decision tree classifier for the .csv file of our training dataset with tree max depth = 5.

TABLE 3: A sample of the training dataset to build a predictive model

Patient State	Symptom name	*TD*	*SSC*	*SAT*	*SP*	Anemia	Asthma	Gastric	Gastric ulcer	Hypertension	Migraine	Myocardial Infraction	Pneumonia	Kental Stone
Person Fifteen	Headache	0:19:19	1	19	34.55	3.455	0	0	0	10.365	20.73	0	0	0
Person Fifteen	Vomiting	1:12:11	2	36	65.45	0	0	32.725	22.9075	0	9.8175	0	0	0
Person Nine	Headache	3:05:03	6	31	100	10	0	0	0	30	60	0	0	0
Person Fourty_ Five	Abdominal_ Pain	2:44:42	5	33	44.59	0	0	0	26.754	0	0	0	0	17.836
Person Fourty_ Five	Headache	2:47:45	6	28	37.84	3.784	0	0	0	11.352	22.704	0	0	0
Person Fourty_ Five	Vomiting	0:51:51	4	13	17.57	0	0	8.785	6.1495	0	2.6355	0	0	0

3. Samples are the number of observations in a node.
4. Values are the number of samples for each node.

We have four network metrics, namely, *TD*, *SSC*, *SAT*, and *SP* as feature variables and the symptoms as target variables; therefore, the data contains five samples for each node. Furthermore, the leaf nodes do not have a question, since they represent the last predictions that are made.

In addition, cross-validation is an important technique, which allows us to use our data better Breiman [40]. The primary goal of the cross-validation is to evaluate how the results of a model will generalize to an autonomous dataset and to limit problems such as overfitting and underfitting [12]. In our system, simple K-folds strategy is applied to split data into K number of splits, which are called Folds. K-Folds are only applied on DTC, which splits the training data into K parts, as shown in Fig. 9.

5.2. KNN

In this paper, we used Euclidean distance, and it should be noted that the Euclidean distance measures are only acceptable for continuous variables [41], as the data in this project's dataset. To calculate Euclidean distance, the below formula is used.

$$\text{Euclidean distance}(\text{d}(\text{x},\ \text{y})) = \sqrt{\sum_{i=1}^{k}\left(x_i - y_i\right)^2} \quad (6)$$

Where *xi* and *yi* are the variables of vectors *x* and *y*, respectively, in the two-dimensional vector space.

The reason of choosing Euclidean distance to calculate the distance between two data points is that the type of the data of this project is continuous data, and according to Kumar [29] and Huang *et al.* [41] it is one of the fastest methods to find the distance between two points. The value of K is crucial; therefore, a large value provides more precision since it reduces the noise, even though there is no guarantee. In our system, the prediction model gets better results for pre-diagnosing diseases with K = 3, as shown in Fig. 10.

5.3. SVM

In this study, multiclass SVM is applied to both training and test datasets to build a predictive model that is used to analyze the symptoms and diagnose diseases. Thus, SVM algorithm is applied on training data to create a model that allocates all linear samples into five different categories.

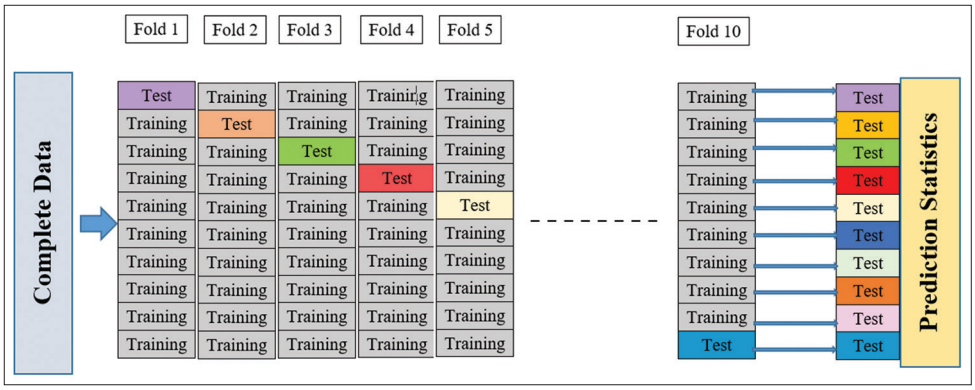

Fig. 9. Cross-validation technique with 10 K-folds.

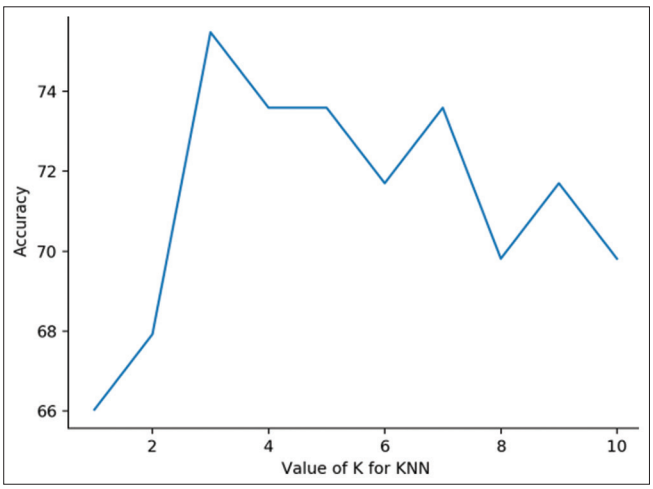

Fig. 10. K values with their accuracy results using K-nearest neighbor algorithm.

Then, new samples in the test data are assigned to a category that is already classified, and it is predicted. Fig. 11 depicts the visualization of SVM for the training and test data. The instances of five train classes and test classes are marked with red, green, blue, black, and yellow, and various shapes that represent different types of symptoms. As it can be seen, the SVM visualization of the training data contains more data than the SVM visualization of the test data, since the overall data are divided such that the part used for training represents 75%, and the part used for testing is 25%.

6. EXPERIMENTAL RESULTS

The system is trained using five types of gestures. Each gesture represents one type of symptom, as shown in Fig. 2. These gestures are read directly from the Sign-Language-Digits-Dataset. The training part uses three different machine learning algorithms: DTC, KNN, and SVM.

Network metrics are computed by applying the mathematical equations that are presented in section 5.3, and then they are stored in *.sv* files. Hence, after training the model, different *.csv* files are used to test the predicted model since the test file is different from the training file, but they contain the same network metrics. The reason why the test data have been separated from the training data is that the predicted model that has been built by the training data did not see the test data. Therefore, according to Dobbin and Simon [42] and Shafique and Hato [43], the result of the machine learning algorithms would be more accurate, reliable, and trustworthy, because they have been tested on new data for the prediction model.

We conducted our experiments using the above-mentioned three machine learning techniques. In addition, there must be measurements for the accuracy of the reorganization performance of the proposed system. One of the most valuable measurement units is the confusion matrix that can be used for recognizing the number of correctly and incorrectly classified instances [44]. The following are the main components of the confusion matrix, and the system uses them to classify the number of correct and incorrect symptoms.

- True positive (*TP*) – Is the classification of an instance with positive class value as a positive case, which means there is a type of symptom and the system predicts correctly.
- False negative (*FN*) – Is the classification of an instance with positive class value as a negative case, which means there is not any symptom and the system does not predict as well.
- False positive (*FP*) – Is the classification of an instance with negative class value as a positive case, which means there is not any symptom and the system is predicting.
- True negative (*TN*) – Is the classification of an instance with negative class value as a negative case, which means there is a symptom, but the system does not predict.

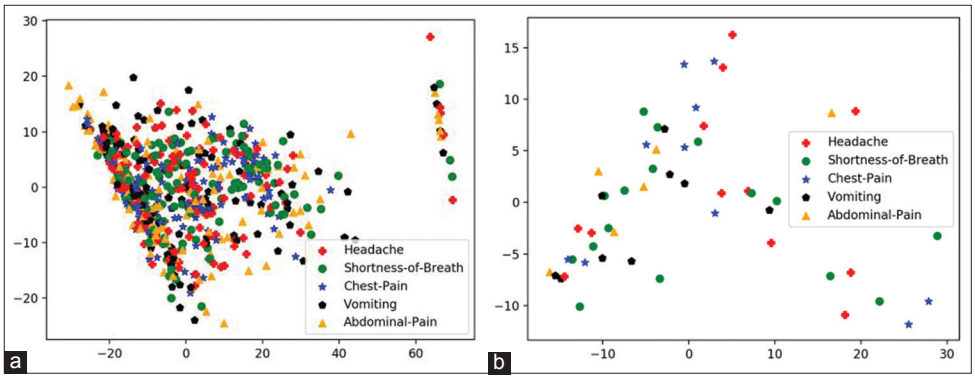

Fig. 11. Visualization of the support vector machine (SVM) for the .csv files of the training and test datasets. (a) Visualizing SVM for training data (b) visualizing SVM for testing data.

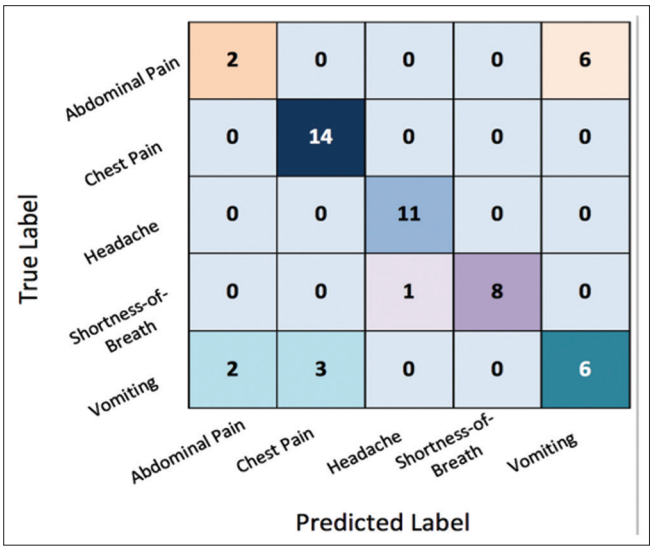

Fig. 12. Confusion matrix results for all types of symptoms.

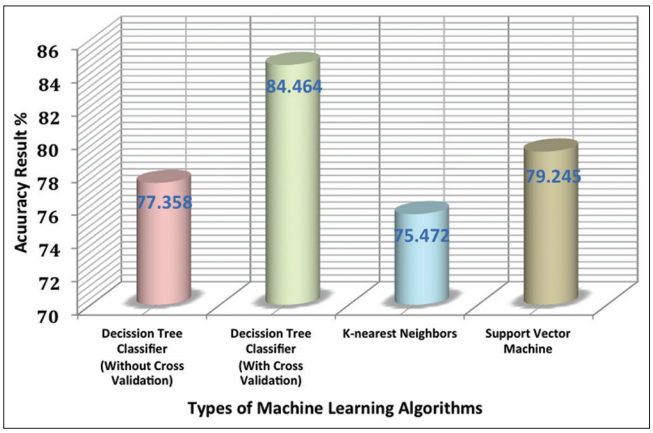

Fig. 13. Accuracy results of all types of machine learning algorithms.

TABLE 4: Confusion matrix classification report

Symptom	Precision	Recall	F1-score	Support
Abdominal pain	0.50	0.25	0.33	8
Chest pain	0.82	1.00	0.90	14
Headache	0.92	1.00	0.96	11
Shortness of breath	1.00	0.89	0.94	9
Vomiting	0.50	0.55	0.52	11

To achieve the most accurate results with the highest performance, the values of *TP* and *FN* must be increasing, and the values of *FP* and *TN* must be decreasing. As a result, the proposed system tries to increase the number of *TP* and *FN* to obtain accurate values. The equation below is used to find the accuracy using the confusion matrix for all three different types of machine learning techniques.

$$(TP/(TP + FN)) * 100 \quad (7)$$

Fig. 12 shows the experimental results in the form of a confusion matrix. It represents all possibilities of the four main components of the confusion matrix with symptoms.

The confusion matrix illustrates how many types of symptoms are correctly classified against misclassified ones. Consequently, the system uses a classification report to check the quality of the predictions of the classification algorithm. A simple example of a classification report is shown in Table 4.

The report explains the major classification metrics, which are precision, recall, F1-score, and support computed from the test dataset. These metrics are defined depending on the four confusion matrix components, which are: TP, TN,

TABLE 5: Comparison between the accuracy of the proposed system with the accuracy of the prior studies

S. No.	Authors	Data provenance	Machine learning algorithms	Accuracy results (%)
1.	Choudhury and Gupta (2019), [14]	NO	Logistic regression	77.61
			Naive Bayes	76.64
			SVM	75.68
			KNN	75.10
			DTC	67.57
2.	Enriko (2019), [31]	NO	Random forest	78.0
			KNN	71.6
			MLP	63.8
			Naive Bayes	50.4
			SVM	45.1
3.	Lavanya *et al*. (2019), [32]	NO	Ada boost	65.50
			SVM	62.93
			DTC	57.75
4.	Alaoui *et al*. (2019), [33]	NO	Neural network	76.73
			SVM	75.43
			Naive Bayes	74.23
			DTC	69.83
5.	Our research	YES	DTC	84.463
			SVM	79.243
			KNN	75.473

KNN: K-nearest neighbor, DTC: Decision tree classifier, SVM: Support vector machine

FP, and FN. From this confusion matrix, we have calculated accuracy for DTC, KNN, and SVM techniques. The accuracy percentages of DTC with and without using cross-validation are 84.464% and 77.358%, respectively. For KNN, the accuracy is 75.472%, and for SVM, it is 79.245%, as shown in Fig. 13. Out of the three techniques the DTC with cross-validation provides the best accuracy.

The accuracy of the results recorded by the proposed system is compared with four prior studies, which were reviewed in section related works. According to these comparisons, our accuracy results are higher than the previous results, as presented in Table 5. The reason for this improvement is the usage of data provenance since it helps the system to record patient's information in timely fashion and records the relations between entities, activities, and agents that would aid the system to make better decisions. Moreover, data provenance simplifies the complexity of huge datasets because of graph representation. Therefore, the novelty of our research is applying data provenance with machine learning algorithms to diagnose a disease.

7. CONCLUSIONS AND FUTURE WORK

Innovative HCI technologies, like gesture recognition, have a great potential for providing and improving the usability and accessibility in interactive systems for aging or disabled people. Moreover, electronic health record systems may incorporate a great deal of useful information about the history of patients' health conditions, which can be shared with clinicians as pre-diagnostic systems.

In this paper, a system is built to pre-diagnose diseases. It starts with the process of hand gesture recognition. Patients use hand gestures to interact with the system. Each gesture represents a symptom. After that, we determine the relations among symptoms and diseases, since symptoms could influence or be influenced by diseases. The symptoms and their relation to various diseases are recorded and stored as provenance graphs. However, our system does not suggest any examination or test for the patients or recommend any medications, rather it decides on which diseases are related to the symptoms by providing percentages that show the mutual

effects between symptoms and diseases. This information can help clinicians make successful decisions in three main terms: Effectiveness-in terms of selecting the least number of correct actions required per view, safety-in term of decreasing the number of incorrect actions, and productiveness-in term of implementing the tasks successfully and efficiently in less time.

Several case studies can be conducted in the future. We will build hand gesture recognition in a timely fashion to improve patient care in real-time, rather than simply recording past events. This will assist physicians to achieve tasks more efficiently and productively. In addition, our system uses provenance graphs to represent patients' health situations and comprehending provenance information would be a challenge to someone who is not an expert in computing technologies. Therefore, it is crucial to convert provenance graphs to textual explanations that could help clinicians to better understand the patient's situation and make more accurate decisions.

8. ACKNOWLEDGMENTS

We would like to express our special thanks and gratitude to Dr. Sdiq N. Abu-Bakr and Dr. Sahand M. Ali (Sulaimani 400 Beds (Shar) Hospital), and Dr. Redwan B. Kaka-Bra (Sulaimani General Hospital/Emergency) for their valuable information and clarifications on the relations between symptoms and diseases. In addition, based on their experiences, they helped in specifying a suitable ratio of disease.

REFERENCES

[1] T. Ganokratanaa and S. Pumrin. The vision-based hand gesture recognition using blob analysis. *International Conference on Digital Arts, Media and Technology, Institute of Electrical and Electronics Engineers*, pp. 336-341, 2017.

[2] Y. Yang. Gesture Controlled User Interface for Elderly People, *MSc Thesis*, College of Applied Sciences, Oslo and Akershus University, 2016.

[3] World Health Organization. *World Report on Disability*. Geneva, World Health Organization, 2011.

[4] W. Chen. Gesture-based applications for elderly people. *In International Conference on Human-Computer Interaction*, Springer, Berlin, Heidelberg, 2013, pp. 186-195.

[5] R. Orji and K. Moffatt. Persuasive technology for health and wellness: State-of-the-art and emerging trends. *Health Informatics Journal*, vol. 24, no. 1, pp. 66-91, 2018.

[6] S. Nowozin, P. Kohli and J. D. Shotton. Gesture Detection and Recognition." *U.S. Patent No. 9,619,035*, 2017.

[7] O. Asan and E. Montague. Technology-mediated information sharing between patients and clinicians in primary care encounters. *Behaviour and Information Technology*, vol. 33, no. 3, pp. 259-270, 2014.

[8] H. Kaur and J. Rani. A review: Study of various techniques of hand gesture recognition. *In 2016 IEEE 1st International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES)*, pp. 1-5, 2016.

[9] D. V. Froy and F. Idris. *Continuous Gesture Recognition for Gaming Systems*," U.S. Patent Application No 10/290, I. G. T. Inc, 2019, p. 176.

[10] L. Moreau and P. Missier. PROV-DM: The PROV data model. *Tech. Rep. W3C Recommendation*, W3C: Available from: http://www.w3.org/TR/prov- dm. 2013. [Last accessed on 2019 Apr 18].

[11] S. Xu, T. Rogers, E. Fairweather, A. Glenn, J. Curran and V. Curcin. Application of data provenance in healthcare analytics software: information visualisation of user activities. *AMIA Summits on Translational Science Proceedings*, vol. 2018, pp. 263-272, 2018.

[12] T. D. Huynh, M. Ebden, J. Fischer, S. Roberts and L. Morea. Provenance network analytics. *Data Mining and Knowledge Discovery*, vol. 32, no. 3, pp.708-735, 2018.

[13] B. Shickel, P. J. Tighe, A. Bihorac and P. Rashidi. Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 5, pp. 1589-1604, 2017.

[14] A. Choudhury and D. Gupta. A survey on medical diagnosis of diabetes using machine learning techniques. *In Recent Developments in Machine Learning and Data Analytics*, Springer, Singapore, 2019, pp. 67-78.

[15] B. Roper, A. Chapman, D. Martin and J. Morley. A graph testing framework for provenance network analytics. *In International Provenance and Annotation Workshop*, Springer, Cham, 2018, pp. 245-251.

[16] T. Lebo, S. Sahoo and D. McGuinness. PROV- O: The PROV Ontology, 2013. Available from: http://www.w3.org/TR/prov-o. [Accessed: 10-April-2019].

[17] H. Miao and A. Deshpande. Understanding data science lifecycle provenance via graph segmentation and summarization, *arXiv preprint arXiv: 1810.04599*, 2018.

[18] G. Closa, J. Maso, B. Proß and X. Pons. W3C prov to describe provenance at the dataset, feature and attribute levels in a distributed environment. *Computers, Environment and Urban Systems*, vol. 64, no. 1, pp. 103-117, 2017.

[19] J. Cheney, P. Missier, L. Moreau, T. DeNies. Constraints of the PROV data model, *Tech. Rep., W3C Recommendation*, W3C: Available from: http://www.w3.org/TR/prov-constraints. 2013. [Last accessed on 2019 Apr 18].

[20] L. Moreau, T. D. Huynh and D. Michaelides. An online validator for provenance: Algorithmic design, testing, and API, *In International Conference on Fundamental Approaches to Software Engineering*, Springer, Berlin, Heidelberg, 2014, pp. 291-305.

[21] J. Hussein, L. Moreau and V. Sassone. Obscuring provenance confidential information via graph transformation, *IFIP Advances in Information and Communication Technology*, vol. 454, ISBN 978-3-319-18491-3, Springer International Publishing, 2015, pp. 109-125.

[22] J. Hussein and L. Moreau. A template-based graph transformation system for the PROV data model. *In Seventh International Workshop on Graph Computation Models*, pp. 1-15. 2016.

[23] L. Moreau and P. Groth. Provenance: An introduction to PROV. *Synthesis Lectures on the Semantic Web: Theory and Technology*, vol. 3, no. 4, pp. 1-129, 2013.

[24] I. Portugal, P. Alencar and D. Cowan. The use of machine learning algorithms in recommender systems: A systematic review. *Expert*

*Systems with Applications, 9*7, pp. 205-227, 2018.

[25] M. Fatima and M. Pasha. Survey of machine learning algorithms for disease diagnostic. *Journal of Intelligent Learning Systems and Applications*, vol. 9, no. 1, pp. 1-16, 2017.

[26] B. L. Deekshatulu and P. Chandra. Classification of heart disease using k-Nearest neighbor and genetic algorithm. *Procedia Technology*, vol. 10, no. 3, pp. 85-94, 2013.

[27] B. J. Erickson, P. Korfiatis, Z. Akkus and T. L. Kline. Machine learning for medical imaging. *Radiographics*, vol. 37, no. 2, pp.505-515, 2017.

[28] A. Yahyaoui and N. Yumusak. Decision support system based on the support vector machines and the adaptive support vector machines algorithm for solving chest disease diagnosis problems. *Biomedical Research*, vol. 29, no. 7, pp. 1474-1480, 2018.

[29] S. Kumar. Activity recognition in egocentric video using SVM, KNN and combined SVM and KNN classifiers, *In IOP Conference Series: Materials Science and Engineering*, vol. 225, no. 1, IOP Publishing, 2017, p. 12226.

[30] M. Pereda and E. Estrada. Machine learning analysis of complex networks in Hyperspherical space. *arXiv preprint arXiv: 1804.05960*, 2018.

[31] I. Enriko. Comparative study of heart disease diagnosis using top ten data mining classification algorithms, *In Proceedings of the 5th International Conference on Frontiers of Educational Technologies*, 2019, pp. 159-164.

[32] B. M. Lavanya, B. H. Kishore, R. Sreekala and S. Sneha. Diagnosis of Indian patients liver disease using machine learning techniques. *International Journal of Research in Advent Technology*, vol. 7, no. 4, pp. 389-391, 2019.

[33] S. S. Alaoui, B. Aksasse and Y. Farhaoui. Data Mining and Machine Learning Approaches and Technologies for Diagnosing Diabetes in Women. *In International Conference on Big Data and Networks Technologies*, Springer, Switzerland, 2019, pp. 59-72.

[34] S. Patel and G. Desai. A comparative study on data mining tools. *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 4, no. 2, pp. 28-30, 2015.

[35] A. Haria, A. Subramanian, N. Asokkumar, S. Poddar and S. Nayak. Hand gesture recognition for human computer interaction. *Procedia Computer Science*, vol. 115, pp. 367-374, 2017.

[36] D. Anderson. *Medical Terminology: The Best and Most Effective Way to Memorize, Pronounce and Understand Medical Terms*. 2nd ed. Independently Published, US, 2016.

[37] F. Luus, N. Khan, and I. Akhalwaya. Active Learning with Tensor Board Projector, *arXiv preprint arXiv: 1901.00675*. January 2019.

[38] F. Beser, A. Kizrak, B. Bolat and T. Yildirim. Recognition of sign language using capsule networks, *In 2018 26th Signal Processing and Communications Applications Conference (SIU)*, IEEE, 2018, pp. 1-4.

[39] M. Arda and D. Zeynep. Sign-language-digits-dataset. *Kaggle Dataset*, vol. 9, pp. 17-25, 2017.

[40] L. Breiman. *Classification and Regression Trees*. New York, US, John Wiley and Sons, Inc, 2017.

[41] G. Huang, C. Guo, M. J. Kusner, Y. Sun, F. Sha and K. O. Weinberger. Supervised Word Mover's Distance. *In Advances in Neural Information Processing Systems*, 2016, pp. 4862-4870.

[42] K. Dobbin and R. M. Simon. Optimally splitting cases for training and testing high dimensional classifiers. *BMC Medical Genomics*, vol. 4, no. 1, pp. 31-39, 2011.

[43] A. Shafique and E. Hato. Formation of training and testing datasets, for transportation mode identification. *Journal of Traffic and Logistics Engineering*, vol. 3, no. 1, pp. 77-80, 2015.

[44] K. M. Ting. Confusion matrix. *Encyclopedia of Machine Learning and Data Mining*, Springer, Boston, MA, 2017, pp. 260-260.