Heart Disease Classification for Early Diagnosis based on Adaptive Hoeffding Tree Algorithm in IoMT Data

Ersin Elbasi
College of Engineering and Technology
American University of the Middle East, Kuwait
ersin.elbasi@aum.edu.kw

Aymen I. Zreikat
College of Engineering and Technology
American University of the Middle East, Kuwait
aymen.zreikat@aum.edu.kw

Abstract: *Heart disease is a rapidly increasing disease that causes death worldwide. Therefore, scientists around the globe start studying this issue from a different perspective to assure early prediction of diagnosis to save patients' life from bad consequences that cause death. In this regard, Internet of Medical Things (IoMT) applications and algorithms should be utilized effectively to overcome this problem. Hoeffding Tree Algorithm (HTA) is a standard decision tree algorithm to handle large sizes of data sets. In this paper, an Adaptive Hoeffding Tree (AHT) algorithm is suggested to carry out classifications of data sets for early diagnosis of heart disease-related factors, and the obtained results by this algorithm are compared with other suggested Machine Learning (ML) algorithms in the literature. Therefore, a total of 3000 records of data sets are used in the classification, 33% of the data are utilized for female patient information, and the rest of the data are utilized for male patient information. In the original data set, each patient record includes 76 attributes, however only the most important 16 patient attributes are used for the classification. Data are retrieved from the University of California Irvine (UCI) Machine Learning Repository, which is collected from the Hungarian Institute of Cardiology, University Hospital at Zurich, University Hospital at Basel, and V.A. Medical Center. The obtained results from this study and the provided comparative results show the effectiveness of the AHT algorithm over other ML algorithms. Compared to other ML algorithms, AHT outperforms other algorithms with 95.67% accuracy for early estimation of diagnosis of heart disease.*

Keywords: *Internet of medical things, machine learning, medical data, random forest, internet of things, diagnosis, AHT.*

Received August 7, 2021; accepted September 26, 2022
https://doi.org/10.34028/iajit/20/1/5

1. Introduction

Nowadays, heart diseases or specifically heart attacks are affecting a broad range of patients in the community around the world. As stated by the World Health Organization (WHO), Cardiovascular Diseases (CVDs) are the main reason for the deaths of about 17.9 million people. Most of the given diagnoses presented to patients are based on the doctor's knowledge only, and therefore they are subject to mistakes from time to time. Consequently, researchers start looking at new techniques or algorithms for the early prediction and detection of Cardiovascular Disease (CVD) as the main cause of heart attack.

To serve humanity's life, the Internet of Things (IoT) and the Internet of Medical Things (IoMT) technologies are utilized by integrating different physical and technological components to work together with other devices over the internet, such as sensors and software. Due to the influence of different technologies such as; embedded systems, wireless sensor networks, control systems, and machine learning [14], IoT applications are growing up worldwide. Currently, the health care industry and applications are completely affected by the quick growth of the IoTs applications [31].

Conventionally, to diagnose the medical conditions of patients in the medical industry, traditional images and scanners are utilized. Therefore, to serve the patients and guarantee better service for them, cutting-edge techniques are desirable to support the medical industry sector with new technologies. Besides, smart systems are needed for early detection of symptoms to discover diseases at primary stages and hence better treatment will be provided for patients. Figure 1 demonstrates the data collection process from several IoT sources.

In the medical industry, Machine Learning (ML) technology is a method of data analysis that is used for diagnoses, therapists, and treatment of various diseases. On the other hand, when huge information about patients in the medical industry is needed to be accessed then deep learning algorithms are utilized for analysis and automation. In this regard, deep learning algorithms are combined with the IoMT to deliver better service and advanced technology for accessing massive medical information of patients, providing an accurate diagnosis of diseases, and properly handling medical image processing. Therefore, the medical industry and insurance companies will benefit from this technology so that the medical services directed to the patients will be optimized.

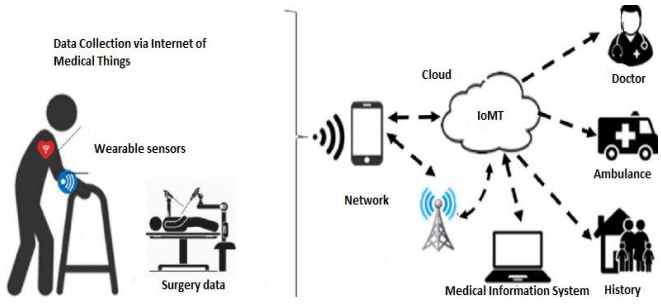


Figure 1. Data streams collected by different internet of medical things sources.

The basic algorithm used in the literature for data stream classification is called Hoeffding Tree (HT) algorithm [24]. It is based on the idea of generating a decision tree that is capable of learning from a set of data streams by assuming that the generated examples are constant over time. Besides, HT algorithm is based on Hoeffding bounds HB, which is independent of the probability distribution assumed in the analysis. HT algorithm has many advantages; it is incremental, provides high accuracy of results even with small samples, and only accepts one scan on the same data. On the other hand, once the node is created using this algorithm, it cannot be eliminated or changed.

HT algorithm will be used in this study to analyse the data set of streams for early prediction and detection of cardiovascular diseases that may cause heart diseases or heart attacks. A modified version of HT is called adaptive size. It is based on the following concept: the tree has a maximum size or split nodes and therefore, if the number of split nodes after the last split node is greater than the maximum value, then the node will be deleted. Specific nodes will be deleted from the tree to satisfy the required threshold for the maximum value. Different techniques to delete some specific nodes from the tree are used in the literature. Either by deleting the last generated node or by resetting the tree and generating the tree nodes from the beginning. In this study, comparative results are provided to show the effectiveness of the AHT algorithm compared to other ML algorithms [14].

2. Literature Review

In the literature, the IoT and more specifically, the IoMT applications in the healthcare industry are discussed and studied intensively and from different approaches. However, many vital aspects should be considered while accessing sensitive information for patients such as reliability, safety, and security [21]. The challenges facing people who are working with IoMTs applications are discussed in [21] by providing a literature review of related research and suggesting a new technique such as a cyber-physical methodology to overcome these challenges. Also, the authors have proposed new research fields to manage most of IoMT's existing challenges. Durga *et al*. [12] have discussed and compared different existing machine learning and deep learning algorithms that are suggested in the literature to serve the healthcare industry. Issues regarding the accuracy and error percentages of these algorithms are discussed by presenting different data sets used in previous research experiments. To monitor the health progress of patients, the authors have proposed a smart IoT-based health care system with a smart medicine box and sensors attached to a centralized database. Using this system, it will be very convenient for doctors to control and observe patients' status in emergency cases, and thus, doctors will be able to follow up with patients' progress, medication, and even interfere. In other research work, authors have proposed in [26] an interactive medical system for the internet of things and cloud platform. The main objective of this system is to improve identity recognition and medical record recognition to enable smooth communication between doctors, patients, and families. A Tri-Storage Failure Recovery System (Tri-SFRS) is proposed in [8] to tackle the problem of processing huge data since single cloud computing is no longer possible to cover enormous data that need to be processed by the healthcare community. As a result of comparing the proposed system with other systems in the literature, the authors proved that the proposed system could improve performance by reducing the latency of data processing.

Due to huge applications regarding IoT and IoMT in the medical industry, it becomes a crucial task to provide patients with a new system to handle their sensitive and private data securely. Therefore, to tackle these issues, Rauscher and Bauer [38] have proposed a healthy and safe approach. In this paper, to cover all possible security and safety issues, the analysis of the Meta-model was established on diverse scenarios. A use case is presented in the evaluation step to validate the suggested architecture for safety and health matters related to IoT medical devices. An adaptive neuro-fuzzy interference and modified salp swarm optimization system is suggested in [29]. This system will help doctors to control heart disease diagnosis. It is shown by authors in this research that it is possible to examine various parameters for a good estimation of heart diseases such as; Blood Pressure (BP), age, sex, chest pain, cholesterol, and blood sugar. By the presented simulation results, the authors claimed that their method provides higher accuracy for predicting heart diseases compared to other studies in the literature. Using deep learning and Adaptive Hybridized Deep Convolutional Neural Networks (AHDCNN), Chen *et al*. [9] have studied the prediction of chronic kidney cancer disease. Based on the suggested scenarios and approaches, the authors have trained different datasets for different kidney layers and the same pattern is tested twice using the deep-belief network to discover early the incorrect pattern that may cause cancer kidney disease. The given statistics and results provide hopeful consequences for this type of disease. The IoMT model is considered to

be the best choice for the effectiveness of predicting diseases. In this respect and to increase the effectiveness of medical diagnosis in IoMT environments, a smart medical service model is suggested in [33]. Based on the presented statistical analysis, the effectiveness of the proposed model in the early discovery of the level of sugar that causes diabetes was verified. Because IoMT is composed of numerous appliances that are connected to provide enhanced health services for patients, crucial factors are considered major challenges and essential issues for IoMT applications such as; storage, processing power, privacy, and security. In [37], the above issues are studied by the authors by proposing a new framework with a privacy and security model that grasps all of the stated issues. However, in [22] other performance parameters are examined to handle diverse IoMT applications such as congestion and delay in the network. An event-aware priority scheduling for data packets to handle the congestion in the network is proposed by authors based on a priority given for emergency packets and therefore delay will be reduced.

Due to technological developments nowadays, there was an increase in generating different data streams based on the interval of time limits such as mobile applications, sensor applications, emails, web applications, and Twitter applications. Heart disease is one of the most common diseases worldwide that need to be studied to better predict the diagnosis of patients before facing a heart attack. Therefore, this issue has been discussed in the literature using different classification algorithms for the obtained data streams.

In [1], a cyber-power Event and Intrusion Detection System (EIDS) is studied for binary class and multi-class classification of power systems and cyber-attack data streams. In this classification, Hoeffding adaptive trees (HAT) with Drift Detection Method (DDM) and Adaptive Windowing (ADWIN) are used. More than 94% accuracy is obtained in multi-class classification, whereas 98% is obtained in a binary class classification. Data science algorithms have been used in [27] to generate a hybrid model to predict heart disease diagnosis in early stages and therefore, suggest a procedure to be considered by patients to avoid any future complications. The provided numerical results in this paper demonstrate a 2% increase in the accuracy of the model. In [28], three machine learning algorithms are suggested to predict the diagnosis of heart disease, namely a hybrid model of random forest and decision tree algorithms have been suggested in this research work. The authors claimed that 88.7% accuracy of prediction is achieved using the suggested techniques. Satu *et al*. [39] claimed that significant factors of heart disease are obtained by studying two significant sets of data for heart disease prediction; namely, Cleveland and Hungarian data set. Heart disease data are studied via semi-supervised learning algorithms and different performance metrics are observed such as accuracy, F-measure, and Area Under ROC (AUROC) to describe the best classifier for the semi-supervised learning algorithm. However, the authors claimed that significant factors of heart disease are observed but they did not give a numerical value of the level of obtained accuracy using this technique. Smart Heart Disease Prediction (SHDP) method with Navies Bayesian algorithm is considered in [32] to predict the diagnosis of heart disease. The authors considered some input parameters to the algorithm that are considered major factors for heart disease such as age, blood sugar, and cholesterol. Part of the data set is used for training and the other part is used for testing. The suggested approach is based on three steps; data collection, patient registration, classification using Navies Bayesian algorithm and finally delivering the produced information to the patient in a secure way using an advanced encryption method. The authors claimed that the suggested approach helped in predicting useful information about heart disease, but they did not provide any measures for the obtained accuracy. A study of circadian Heart Rate Variability (HRV) in CAD patients with various degrees of left ventricle Ejection Fraction (EF) is conducted in [3] to predict heart rate failure among patients. Heart rate variability is studied at different levels; normal level (if EF > 50%), at-risk level (if EF < 40%), and borderline level where EF is between 40% and 50%. The authors claim that the suggested method would help to provide a good prediction of heart failure among CAD patients. Nayak *et al*. [35], study the diagnosis of heart disease at primitive stages to avoid any complications afterward. Therefore, the frequent item mining technique is proposed and Decision tree classification, Naive Bayes classification, Support Vector Machine classification, and k-NN classification are discussed and suggested to predict the diagnosis at early stages so that treatment will be effective.

Hoeffding Trees with N_{min} Adaptation algorithm is suggested in [2, 20]. The authors in this work claimed that adaptive algorithms usually fixed some parameters in the algorithm and therefore, this will lead to redundant parameters that might cause unneeded calculations and power resumption of the algorithm. Therefore, the authors have proposed N_{min} parameters that eliminate those unnecessary calculations while keeping the accuracy at acceptable levels. Comparative results show that a Very Fast Decision Tree (VFDT) algorithm outperforms concerning consumption rate and accuracy. Transfer-based Hoeffding Adaptive Tree (THAT) method is proposed in [34] to synchrophasor signatures. The provided testing results show that THAT outperforms other algorithms such as Ozabag for computational time and accuracy. Antony and Varghese [6], suggested a hardware accelerator for the Hoeffding tree algorithm with an adaptive naive Bayes predictor in the leaves. Minimum hardware resources are utilized in the implementation while mixed data sets of nominal and numeric values are considered. Authors claim that the implemented system is faster than StreamDm(C++)

system for the same data set. A cyber-power Event and Intrusion Detection System (EIDS) to handle multiclass or binary-class classification is proposed in [1]. Improved Hoeffding Adaptive Trees (HAT) with ADWIN is suggested in the classification of the data set. Based on the statistical results, it is shown that the improved HAT method with DDM and ADWIN improves the accuracy by 94% for the multiclass data set and 98% for binary class sets.

Regarding the proper selection of machine learning approaches to handle the classifications, recently some researchers in the literature use new methodologies to gain more accurate result. For example, in [5] eight different machine learning classification approaches are used to gain better results with promising accuracies. The used approaches are Decision Tree (DT), Multilayer Perceptron (MLP), Logistic Regression (LR), Support Vector Machines (SVM), Naive Bayes (NB), k-Nearest Neighbor (k-NN), Adaboost, and XGBoost (XGB). In other work [36], the K-Nearest technique is used with a value k of 7 based on data mining techniques to predict and classify the data set based on the neighborhood. The authors claim that they were able to gain 77% accuracy with this value of k. Using a hybrid data set is another approach proposed by the authors in [30] by collecting different data sets to be used for classification. According to the obtained results, the authors claimed that better accuracy is gained when the data sets are separated part for testing and part for training. In [21], the authors claimed that with proper validation of the results, an accuracy of around 90.16% is obtained by conducting a classification using a Random Forest ML algorithm based on 303 patients with 14 physiological data sets provided by the Cleveland heart disease lab center.

3. Methodology

There are several methods in the literature for medical data prediction, classification, and clustering. Machine learning algorithms are very powerful methods to classify huge data volumes to provide automatically learning system. ML algorithms can be used for the classification, prediction, and clustering of medical data for diagnosis and treatment. There are several machine learning algorithms proposed in the literature for patient data classification. ML algorithms are categorized as supervised learning, unsupervised learning, and reinforcement learning. In supervised learning, the system learns from past and labelled data to predict future events. On the other hand, supervised learning systems use unlabelled data to infer a function without finding outputs. In addition to these methods, there are several works have been developed for semi-supervised learning and reinforcement learning. Nowadays, medical data collection, processing, storing, and knowledge extraction is an important research area, especially for data that is collected from the internet of things. Machine learning and other AI applications are used in healthcare for efficient diagnosis and treatment [17].

In this research work, several machine-learning algorithms are used in the classification of patient heart disease risk. Some of the algorithms are rule-based such as decision tree, random forest, Hoeffding tree, and adaptive Hoeffding tree algorithms. The random forest algorithm is very flexible, and it gives very efficient results most of the time for medical data processing. RF algorithm is used for both regression and classification purposes [11, 13]. It produces several decision trees, then merges them using the bagging classifier method. RF algorithm-based classification pseudocode is given below [19, 23].

Algorithm 1. Random forest algorithm

INPUT: Given data set D
FOR i=1 to N DO
    Training data D sampling (randomly)
    Create root node R
    Call Build (R)
END
BUILD (R)
IF R contains elements of one single class
  THEN return
ELSE
  Randomly select splitting attributes in R
  Select attribute A with the highest gain value
  Create child notes such as $R_1, R_2, \ldots, R_N$, *etc.*
  FOR i=1 to A DO
    Set the contents of R_i *for all data sets*
    Call function BUILD (R)
  END; END
OUTPUT: Tree model

Another rule-based classification method is the Hoeffding tree algorithm. HT is a very fast decision tree algorithm that does not reuse instances, instead, the HT algorithm waits for new instances. HT algorithm solves several problems in streaming data where instead of reusing instances, it waits for new instances. AHT algorithm compares features better than other classification algorithms. Adaptive Hoffding Tree (AHT) algorithm has less memory consumption and enhance the utilization. On the other hand, AHT algorithm is time costly when there is ties between attributes. In that case of AHT algorithm, if data has noisy outliers, it does not give good classification results. Hoffding Tree algorithm has the following additional improvements:

- Best attributes are calculated to split the data when a specific number of new instances arrives. Normally HT algorithm calculates whenever a new instance comes. This method increases the learning time.
- HT algorithm omits the least promising nodes to save from memory location.
- HT algorithm can start any built decision tree.
- If the gain factor is the same for two attributes, the HT algorithm splits attributes that gain is less than

the threshold.

The AHT algorithm uses a change detector and error estimator which increases the efficiency of the HT algorithm. When change is detected, new trees are created in the AHT algorithm without waiting for new instances to arrive. In this way, the AHT algorithm creates a new tree and extracts rules faster than other rule-based methods. When the newly created tree is more accurate it replaces it with the old tree in AHT. HT bound is calculated as follows [25]:

$$b = \sqrt{\frac{R^2 x ln\left(\frac{1}{p}\right)}{2xN}} \quad (1)$$

Where b is the HT bound, R is the range, N demonstrates the number of observations. HT bound is (1-p); the true mean of the variable is calculated as the ($mean\ value\ of\ N$) $- b$

Algorithm 2. Adaptive Hoeffding Tree algorithm

INPUT: Data set D
Assign HT a tree with a single leaf
FOR i=1 to D do
Sort all data into leaf using HT algorithm
Update leaf nodes
Increment each n which is the element of the leaf
IF ni mod nmin =0 and leaf elements are not in the same class
THEN Calculate Gl for each attribute
Assign Dh to be the highest Gi
Assign Ds to be second-highest Gi
Calculate Hoeffding bound ε
IF Dh ≠ D0 and (Gl (Dh)-Gl(Ds) > ε)
THEN replace leaf node with internal node
FOR all branches
Add a new leaf
END
END
END; END
OUTPUT: Adaptive Hoeffding Tree model

4. Experimental Results

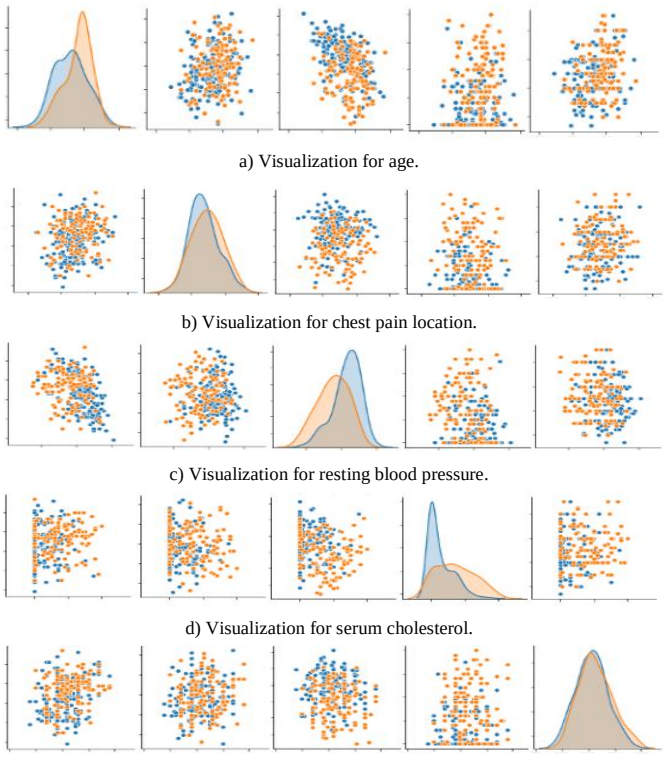
In this work, a total of 3000 patient records were used for prediction and classification. Table 1 demonstrates minimum, maximum, mean, and standard deviation values of some of the attributes such as Chest Pain Type (CPT), Resting Blood Pressure (RBP), serum cholesterol in mg/dl (SC), ST depression induced by exercise related to rest (STD), the slope of the peak exercise ST segment (slope), number of major vessels coloured by fluoroscopy (vessel), thalassemia (T) and patient age. In this work, Weka data mining tool is used in all experiments for classification including random forest classifier and adaptive Hoeffding Tree algorithms.

Table 1. Characteristics of input data CPT, RBP, SC, STD, Slope, V, and T.

Attribute	Minimum	Maximum	Mean	Standard Deviation
CPT	0	3	0.959	1.035
RBP	94	200	131.551	17.53
SC	126	564	246.808	51.893
STD	71	202	149.709	22.982
Slope	0	6.2	1.034	1.147
Vessel	0	2	1.401	0.615
T	0	3	2.315	0.612

In addition to Table 1, other data characteristics are given in below:

- Gender={998 females, 2002 male}
- Age={29-77}
- Fasting blood sugar={2490 normal, 510 abnormal}
- Resting electrocardiographic results={1460 normal, 1540 abnormal}
- Exercise induced angina={1950 yes, 1050 no}
- Number of major vessels colored by fluoroscopy={0,1,2,3,4}



a) Visualization for age.

b) Visualization for chest pain location.

c) Visualization for resting blood pressure.

d) Visualization for serum cholesterol.

e) Visualization for resting electrocardiographic.

Figure 2. Feature vector data visualization for classification.

Data visualization is given in Figure 2 for a. age, b. chest pain location, c. resting blood pressure, d. serum cholesterol and e. resting electrocardiographic results. Several ML algorithms are used for the prediction and classification of heart disease such as the Bayes network, naïve Bayes classifier, logistic, multiplayer perception, stochastic gradient descent, simple logistic, etc., Table 2 demonstrates the accuracy and build time for each applied algorithm. Adaptive Hoeffding Tree gives 95.67% accuracy, random forest 92.89%, and Hoeffding Tree algorithm gives 92.18% accuracy. Most of the other well-known algorithms give more than 80%

accuracy. The Multilayer perception algorithm needs 0.63 seconds of build time for training data, and the adaptive Hoeffding tree algorithm needs only 0.01 seconds of build time. Experimental results show that the adaptive Hoeffding tree algorithm gives more promising results than other rule-based and probabilistic-based algorithms for classification of the heart data which is collected from the internet of medical things devices [10, 18]. Random forest algorithm handles large and complex dataset. It provides higher accuracy than decision tree algorithms. Hoeffding tree learns from massive data streams. It is efficient and adaptive to use. Adaptive Hoeffding tree algorithms replace current tree with alternative trees if better accuracy is obtained. RF and AHT algorithms have high performance in build and testing times. MP algorithm has very high training time because of the number of iterations. On the other hand, it is very suitable and gives efficient results especially complex, large and limited data sets. Our experiments show that MP algorithm has 85.91% accuracy and 0.63 seconds building time performance. Probabilistic based algorithms such as NBC has 89.32% accuracy and 0.02 seconds build time.

Table 2. Accuracy of each machine learning algorithm and build time.

Algorithm	Accuracy	Build time
Bayes Network (BN)	88.47	0.01
Naïve Bayes Classifier (NBC)	89.32	0.01
Naïve Bayes Updatable (NBU)	89.61	0.02
Logistic (L)	86.76	0.01
Multilayer Perception (MP)	85.91	0.63
Stochastic gradient descent (SGD)	87.61	0.01
Simple Logistic (SL)	85.05	0.57
Voted Perception (VP)	72.37	0.02
Locally Weighted Learning (LWL)	83.06	0.25
Regression (R)	92.74	0.01
Filtered Classification (FC)	86.05	0.03
Iterative Classifier Optimizer (ICO)	91.89	0.01
Decision Table (DT)	81.06	0.02
PART	84.19	0.01
Hoeffding Tree (HT)	92.18	0.01
J48	89.61	0.02
Random Forest (RF)	92.89	0.05
Random Tree (RT)	81.64	0.01
Adaptive Hoeffding Tree (AHT)	95.67	0.01

Patient heart data indicates two classes in this work, normal and abnormal patients for heart disease. Table 4 shows measurement values after the classification of data using machine-learning algorithms. Table 4 below demonstrates the True Positive Rate (TP), False Positive Rate (FP), precision, recall, f-measure, Matthew's Correlation Coefficient (MCC), Receiver Operating Characteristic (ROC) area, Precision-Recall Curves (PRC) area for each class in all machine-learning algorithm used in this work. For example, if we use the Bayesian network in classification, the true positive rate is 0.834 for normal patient class, and 0.924 for abnormal patient class. Table 4 shows that the adaptive Hoeffding tree algorithm gives the best result if compared to the other algorithms. AHT algorithm has very high values in recognition of abnormal patients; on the other hand, the regression algorithm has better results in normal patient identification. Figure 3 shows accuracy rates and Figure 4 demonstrates build time values for machine learning algorithms.

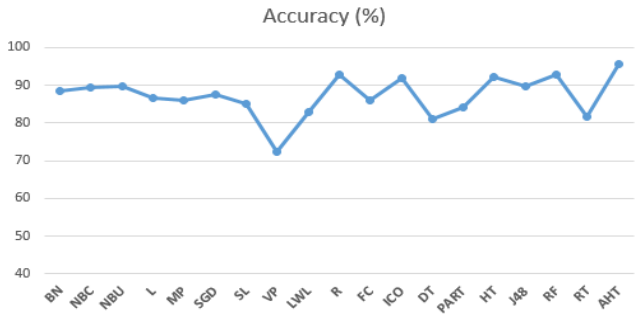


Figure 3. Accuracy rates for ML algorithms.

Figure 5 shows precision, recall, f-measure, Matthew's Correlation Coefficient (MCC), Receiver Operating Characteristic (ROC) area, and Precision-Recall Curves (PRC) area rates for each classification algorithm. Based on the precision graph, FC gives the highest rate and DT gives the lowest rate. VP algorithm has the smallest recall and f-measure values than other algorithms. The proposed AHT algorithm has the highest rate in precision, MCC, f-measure, and recall, on the other hand, a low rate for ROC and PRC areas. It shows the success of the proposed algorithm adaptive Hoeffding tree in the classification of the patient IoMT data [4, 15].

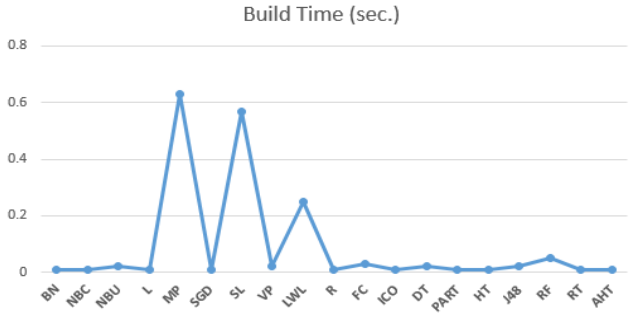


Figure 4. Build time vaues in training for ML algorithms.

Table 3. Kappa, MAE, RMQE, RAE, and RRSE values for ML algorithms.

Algorithm	Kappa	MAE	RMSE	RAE (%)	RRSE (%)
BN	0.7326	0.1368	0.2998	32.53	63.24
NBC	0.7482	0.1387	0.2971	30.92	62.70
NBU	0.7482	0.1387	0.2971	30.92	62.70
L	0.7021	0.1684	0.3134	36.87	65.96
MP	0.6809	0.1404	0.3383	31.25	70.97
SGD	0.7144	0.1238	0.3622	27.92	75.78
SL	0.6675	0.1864	0.3185	40.49	66.99
VP	0.4490	0.2888	0.5114	61.08	92.76
LWL	0.5368	0.2971	0.3751	62.75	78.36
R	0.7544	0.1606	0.2974	35.31	62.75
FC	0.8186	0.2325	0.3134	49.77	65.98
ICO	0.8056	0.1684	0.2908	36.88	61.43
DT	0.5498	0.2899	0.3789	61.31	79.12
PART	0.6510	0.1592	0.3769	35.02	78.73
HT	0.7664	0.1349	0.3003	30.14	63.33
J48	0.7158	0.1863	0.3414	40.48	71.60
RF	0.8025	0.2216	0.2952	47.58	62.32
RT	0.5926	0.1837	0.4323	39.95	89.84
AHT	0.8316	0.1925	0.2641	50.12	63.14

Table 4. TP rate, FP rate, precision, recall, f-measure, MCC, ROC area, PRC area values for ML algorithms.

Method	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
BN	0.834	0.076	0.884	0.834	0.858	0.734	0.947	0.941	N
	0.924	0.166	0.885	0.924	0.904	0.734	0.947	0.950	A
NBC	0.841	0.046	0.919	0.814	0.863	0.752	0.944	0.946	N
	0.954	0.159	0.877	0.954	0.914	0.752	0.944	0.951	A
NBU	0.814	0.046	0.919	0.814	0.873	0.752	0.944	0.946	N
	0.954	0.186	0.877	0.954	0.914	0.752	0.944	0.951	A
L	0.893	0.137	0.826	0.873	0.849	0.703	0.938	0.933	N
	0.863	0.107	0.903	0.863	0.883	0.703	0.938	0.947	A
MP	0.804	0.106	0.846	0.814	0.830	0.681	0.937	0.910	N
	0.894	0.196	0.868	0.894	0.881	0.681	0.937	0.945	A
SGD	0.814	0.076	0.881	0.814	0.846	0.716	0.869	0.792	N
	0.924	0.186	0.873	0.924	0.898	0.716	0.869	0.843	A
SL	0.854	0.152	0.818	0.854	0.830	0.668	0.933	0.928	N
	0.848	0.146	0.887	0.848	0.867	0.668	0.993	0.940	A
VP	0.432	0.061	0.799	0.522	0.539	0.389	0.769	0.730	N
	0.939	0.568	0.689	0.939	0.794	0.389	0.769	0.778	A
LWL	0.697	0.137	0.786	0.697	0.738	0.540	0.906	0.888	N
	0.863	0.303	0.794	0.863	0.837	0.540	0.906	0.932	A
R	0.912	0.122	0.848	0.912	0.879	0.756	0.946	0.919	N
	0.878	0.088	0.933	0.878	0.905	0.756	0.946	0.956	A
FC	0.854	0.015	0.963	0.854	0.905	0.823	0.949	0.937	N
	0.985	0.166	0.905	0.985	0.943	0.823	0.949	0.956	A
ICO	0.795	0.106	0.843	0.795	0.818	0.673	0.874	0.786	N
	0.894	0.205	0.856	0.894	0.874	0.673	0.874	0.878	A
DT	0.814	0.228	0.732	0.814	0.771	0.533	0.857	0.780	N
	0.772	0.186	0.847	0.772	0.818	0.553	0.857	0.873	A
PART	0.854	0.167	0.794	0.854	0.822	0.653	0.866	0.787	N
	0.833	0.166	0.885	0.833	0.858	0.653	0.866	0.852	A
HT	0.834	0.046	0.921	0.834	0.875	0.769	0.946	0.948	N
	0.954	0.166	0.889	0.954	0.921	0.769	0.946	0.951	A
J48	0.834	0.091	0.867	0.834	0.850	0.716	0.870	0.790	N
	0.909	0.166	0.883	0.909	0.896	0.716	0.870	0.861	A
RF	0.873	0.046	0.926	0.873	0.899	0.804	0.963	0.958	N
	0.954	0.127	0.914	0.954	0.934	0.804	0.963	0.961	A
RT	0.755	0.137	0.801	0.755	0.777	0.593	0.809	0.709	N
	0.863	0.245	0.827	0.863	0.845	0.593	0.809	0.788	A
AHT	0.853	0.008	0.941	0.854	0.891	0.816	0.956	0.968	N
	0.992	0.147	0.902	0.982	0.939	0.816	0.956	0.972	A

Table 3 shows error rates for each ML algorithm such as kappa, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Relative Absolute Error (RAE), and Root Relative Squared Error (RRSE). Mean absolute error rate and relative absolute error rates are calculated as follows [6, 16]:

$$MAE = \frac{\sum_{i=1}^{n}|Y_i - X_i|}{N} \quad (2)$$

$$RAE = \frac{\sqrt{\sum_{i=1}^{n}(X-Y)^2}}{\sqrt{\sum_{i=1}^{n}Y_i^2}} \quad (3)$$

Where N is the total number of data, A shows the actual values and B demonstrates expected values. In this work, rule-based classification algorithms give more promising results than other ML algorithms. AHT algorithm creates a total of 1225 rules, after pruning the tree some of the extracted rules are given in Table 5.

In classification algorithms there are three metrics we should consider; the accuracy of algorithm, the needed space and the time complexity. AHT algorithm is very fast, and any case create a classification tree. AHT algorithm compares features better than other classification algorithms. AHT algorithm has less memory consumption and enhance the utilization. On the other hand, AHT algorithm is time costly when there is a ties between attributes. If data has noisy outliers, in that case AHT algorithm does not give good classification results.

Table 5. List of samples extracted rules from AHT.

•IF {(Thalassemia <=2) AND (Vessels colored by fluoroscopy = 0) AND (Exercise induced angina =0)} THEN CLASS: ABNORMAL
•IF {(Thalassemia <=2) AND (Vessels colored by fluoroscopy = 0) AND (Exercise induced angina =1) AND (Resting electrocardiographic results = 1) AND (The slope of the peak exercise ST segment <= 1) AND (Chest pain type > 1)} THEN CLASS: ABNORMAL
•IF {(Thalassemia <=2) AND (Vessels colored by fluoroscopy = 0) AND (Exercise induced angina =1) AND (Resting electrocardiographic results = 1) AND (The slope of the peak exercise ST segment > 1)} THEN CLASS: ABNORMAL
•IF {(Thalassemia <=2) AND (Vessels colored by fluoroscopy > 0) AND (Chest pain type <=0) AND (gender=0) AND (The slope of the peak exercise ST segment <=1)} THEN CLASS: NORMAL
•IF {(Thalassemia <=2) AND (Vessels colored by fluoroscopy > 0) AND (Chest pain type <=0) AND (gender=0) AND (The slope of the peak exercise ST segment >1)} THEN CLASS: ABNORMAL

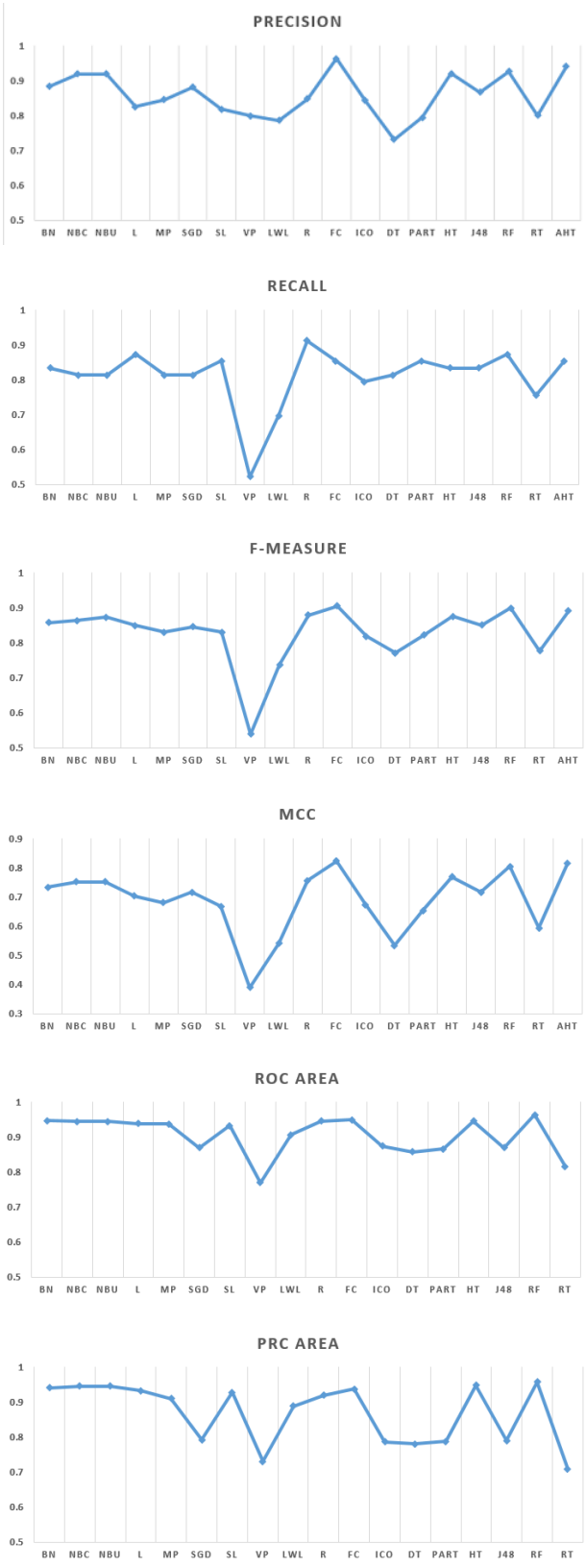


Figure 5. Precision, recall, f-measure, MCC, ROC area, PRC area rates for each classification algorithms.

5. Conclusions, Limitations, and Future Work

Early prediction of heart disease diagnosis will save patients' lives from the bad consequences that cause death to many people around the globe. Therefore, IoMT applications and algorithms should be utilized effectively to overcome this problem. An adaptive Hoeffding tree algorithm is suggested to carry out classifications of data sets for early diagnosis of heart disease-related factors. In this work, a total of 3000 records of data sets are used in the classification, 33% of the data are utilized for female patient information, and the rest of the data are utilized for male patient information. In the original data set, each patient record includes 76 attributes, however only the most important 16 patient attributes are used for classification. Data are retrieved from the UCI machine learning repository [7], which is collected from the Hungarian Institute of Cardiology, University Hospital at Zurich, University Hospital at Basel, and V.A. Medical Center. The obtained results from this study and the provided comparative results show the effectiveness of the AHT algorithm over other ML algorithms. Compared to other ML algorithms, AHT outperforms other algorithms with 95.67% accuracy for early prediction of diagnosis of heart disease. In future work, a new data set will be considered to include other attributes to cover all possible diagnoses that cause heart disease, and hence the probability of cure can be maximized. In this regard and to gain early and better prediction of diagnosis, other sources of data sets should be discovered to cover more properties for the more male and female gender, and hence, more disease-related factors can be exposed with a broader range of diagnoses. For example, the Adaptive Hoeffding tree algorithm is suggested to carry out the classifications in this research work. However, as an extension to the current work, the obtained results should be compared regarding accuracy with other similar work in the literature to show the effectiveness and the accuracy of the obtained results in the early prediction of the disease and therefore, to save the life of more people. Although encouraging results are achieved in this study concerning the accuracy of early prediction of the disease, however, using machine learning algorithms in the classification of data sets for sensitive data such as data related to heart diseases exposes some limitations such as:

- The provided data set does not cover a large range of male and female properties to increase the accuracy of the results. Therefore, discovering new data sets from other resources can be a good future research work.
- The utilized data set in the current study is collected from the Hungarian Institute of Cardiology, University Hospital at Zurich, University Hospital at Basel, and V.A. Medical Center. To cover more diagnoses and provide better predictions of the diseases, the collected data set should be obtained from different areas that are geographically diverse.
- Using hybrid data sets from different resources can be a limitation of this study as some data set resources are not available upon request or are not provided in the English language.

References

[1] Adhikari U., Morris T., and Pan S., "Applying Hoeffding Adaptive Trees for Real-Time Cyber-Power Event and Intrusion Classification," *IEEE Transactions on Smart Grid*, vol. 9, no. 5, pp. 4049-4060, 2018.

[2] Aljanabi M. and Ismail M., "Improved Intrusion Detection Algorithm based on TLBO and GA Algorithms," *The International Arab Journal of Information Technology*, vol. 18, no. 2, pp. 170-179, 2021.

[3] Alkhodari M., Jelinek H., Werghi N., Hadjileontiadis L., and Khandoker A., "Investigating Circadian Heart Rate Variability in Coronary Artery Disease Patients with Various Degrees of Left Ventricle Ejection Fraction," *in Proceedings 42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Montreal, pp. 714-717, 2020.

[4] Al-Mahmud O., Khan K., Roy R., and Mashuque A., "Internet of Things (IoT) Based Smart Health Care Medical Box for Elderly People," *in Proceedings International Conference for Emerging Technology*, Belgaum, pp. 1-6, 2020.

[5] Akkaya B., Sener E., and Gursu C., "A Comparative Study of Heart Disease Prediction Using Machine Learning Techniques," *in Proceedings International Congress on Human-Computer Interaction, Optimization and Robotic Applications*, Ankara, pp. 1-8, 2022.

[6] Antony A. and Varghese K., "High Throughput Hardware for Hoeffding Tree Algorithm with Adaptive Naive Bayes Predictor," *in Proceedings 6th International Conference for Convergence in Technology*, Maharashtra, pp. 1-6, 2021.

[7] Blake C. and Merz C., UCI Repository of Machine Learning Databases, 1998.

[8] Cao R., Tang Z., Liu C., and Veeravalli B., "A Scalable Multi-Cloud Storage Architecture for Cloud-Supported Medical Internet of Things," *IEEE Internet of Things Journal*, vol. 7, no. 3, pp. 1641-1654, 2020.

[9] Chen G., Ding C., Li Y., Hu X., Li X., Ren L., Ding X., Tian P., and Xue W., "Prediction of Chronic Kidney Disease Using Adaptive Hybridized Deep Convolutional Neural Network on the Internet of Medical Things Platform," *IEEE Access*, vol. 8, 2020.

[10] Cihan P. and Coşkun H., "Performance Comparison of Machine Learning Models for Diabetes Prediction," *in Proceedings of Signal Processing and Communications Applications Conference*, Istanbul, pp. 1-4, 2021.

[11] Dhanka S. and Maini S., "Random Forest for Heart Disease Detection: A Classification Approach," *in Proceedings of 2nd International Conference on Electrical Power and Energy Systems*, Bhopal, pp. 1-3, 2021.

[12] Durga S., Nag R., and Daniel E., "Survey on Machine Learning and Deep Learning Algorithms used in Internet of Things (IoT) Healthcare," *in Proceedings of International Conference on Computing Methodologies and Communication* (ICCMC), Erode, 2019.

[13] El-Shafiey M., Hagag A., El-Dahshan E., and Ismail M., "Heart-Disease Prediction Method Using Random Forest and Genetic Algorithms," *in Proceedings of International Conference on Electronic Engineering*, Menouf, pp. 1-6, 2021.

[14] Elbasi E. and Zreikat A., "Efficient Early Prediction and Diagnosis of Diseases Using Machine Learning Algorithms for IoMT Data," *IEEE World AI IoT Congress*, Seattle, pp. 0155-0159, 2021.

[15] Elbasi E., Ayanoglu D., and Zreikat A., "Determination of Patient's Hearing Sensitivity using Data Mining Techniques," *in Proceedings of IEEE 12th International Conference on Application of Information and Communication Technologies*, Almaty, pp. 1-6, 2018.

[16] Elbasi E., "Reliable Abnormal Event Detection from Iot Surveillance Systems," *in Proceedings of 7th International Conference on Internet of Things: Systems, Management, and Security*, Paris, pp. 1-5, 2020.

[17] Elbasi E., Topcu A., and Mathew S., "Prediction of COVID-19 Risk in Public Areas Using IoT and Machine Learning," *Electronics*, vol. 10, no. 14, 2021.

[18] Elbasi E., "B-DCT based Watermarking Algorithm for Patient Data Protection in IoMT," *in Proceedings of International Conference on Information Security and Cryptology*, Ankara, pp. 1-4, 2020.

[19] Fazakis N., Kocsis O., Dritsas E., Alexiou S., Fakotakis N., and Moustakas K., "Machine Learning Tools for Long-term Type 2 Diabetes Risk Prediction," *IEEE Access*, 2021.

[20] García-Martín E., Lavesson N., Grahn H., Casalicchio E. and Boeva V., "Hoeffding Trees with Nmin Adaptation, " *in Proceedings of IEEE 5th International Conference on Data Science and Advanced Analytics*, Turin, pp. 70-79, 2018.

[21] Gatouillat A., Badr Y., Massot B. and Sejdić E., "Internet of Medical Things: A Review of Recent Contributions Dealing with Cyber-Physical Systems in Medicine," *IEEE Internet of Things Journal*, vol. 5, no. 5, pp. 3810-3822, 2018.

[22] Gopikrishnan S., Priakanth P., Srivastava G., and Fortino G., "EWPS: Emergency data communication on the Internet of Medical Things," *IEEE Internet of Things Journal*, 2021.

[23] Han M., Ding J., and Li J., "PatHT: An Efficient Method of Classification over Evolving Data

Streams," *The International Arab Journal of Information Technology*, vol. 16, no. 6, pp. 1098 1105, 2020.

[24] Hulten G., Spencer L., and Domingos P., "Mining Time-Changing Data Streams," *in Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California, pp. 97-106, 2021.

[25] Jayakrishnan A. Visakh R., and Ratheesh T., "Computational Approach for Heart Disease Prediction using Machine Learning," *in Proceedings of International Conference on Communication, Control and Information Sciences*, Idukki, pp. 1-5, 2021.

[26] Jia N. and Zheng C., "Design of Intelligent Medical Interactive System Based on Internet of Things and Cloud Platform," *in Proceedings of International Conference on Intelligent Human-Machine Systems and Cybernetics*, Hangzhou, pp. 28-31, 2018.

[27] Junaid M. and Kumar R., "Data Science and Its Application in Heart Disease Prediction," *in Proceedings of International Conference on Intelligent Engineering and Management*, London, pp. 396-400, 2020.

[28] Kavitha M., Gnaneswar G., Dinesh R., Sai Y., and Suraj R., "Heart Disease Prediction using Hybrid Machine Learning Model," *in Proceedings of 6th International Conference on Inventive Computation Technologies*, Coimbatore, pp. 1329-1333, 2021.

[29] Khan M. and Algarni F., "A Healthcare Monitoring System for the Diagnosis of Heart Disease in the IoMT Cloud Environment Using MSSO-ANFIS," *IEEE Access*, vol. 8, 2020.

[30] Khan M. and Mondal M., "Effectiveness of Data-Driven Diagnosis of Heart Disease," *in Proceedings of International Conference on Electrical and Computer Engineering*, Dhaka, pp. 419-422, 2020.

[31] Laha S., Ghosh D., and Swarnakar B., *Internet of Medical Things for Smart Healthcare*, Springer Singapore, 2020.

[32] Lakshmanarao A., Srisaila A., and Kiran T., "Heart Disease Prediction using Feature Selection and Ensemble Learning Techniques," *in Proceedings of 3rd International Conference on Intelligent Communication Technologies and Virtual Mobile Networks*, Tirunelveli, pp. 994-998, 2021.

[33] Lu S. Wang A., Jing S., Shan T., Zhang X., Guo Y., and Liu Y., "A Study on Service-Oriented Smart Medical Systems Combined with Key Algorithms in The Iot Environment," *China Communications*, vol. 16, no. 9, pp. 235-249, 2019.

[34] Mrabet Z., Selvaraj D., and Ranganathan P., "Adaptive Hoeffding Tree with Transfer Learning for Streaming Synchrophasor Data Sets," *in Proceedings of IEEE International Conference on Big Data (Big Data)*, Los Angeles, pp. 5697-5704, 2019.

[35] Nayak S., Gourisaria M., Pandey M., and Rautaray S., "Prediction of Heart Disease by Mining Frequent Items and Classification Techniques," *in Proceedings of International Conference on Intelligent Computing and Control Systems*, Madurai, pp. 607-611, 2019.

[36] Rahman B., Hendric S., Sabarguna B., and Budiharto W., "Heart Disease Classification Model Using K-Nearest Neighbor Algorithm," *in Proceedings of 6th International Conference on Informatics and Computing*, Jakarta, pp. 1-4, 2021.

[37] Rathnayake R., Karunarathne M., Nafi N., and Gregory M., "Cloud-Enabled Solution for Privacy Concerns in the Internet of Medical Things," *in Proceedings of International Telecommunication Networks and Applications Conference*, Sydney, pp. 1-4, 2018.

[38] Rauscher J. and Bauer B., "Safety and Security Architecture Analyses Framework for the Internet of Things of Medical Devices," *in Proceedings of International Conference on e-Health Networking, Applications and Services*, Ostrava, pp. 1-3, 2018.

[39] Satu M., Tasnim F., Akter T., and Halder S., "Exploring Significant Heart Disease Factors based on Semi-Supervised Learning Algorithms," *in Proceedings of International Conference on Computer, Communication, Chemical, Material and Electronic Engineering*, Rajshahi, pp. 1-4, 2018.

Ersin Elbasi is currently working for American University of the Middle East. He received MSc degree in computer science at Syracuse University; MPhil and PhD degrees in computer science at Graduate Center, The City University of New York. His research interests include multimedia security, event mining in video sequences and medical image processing.

Aymen I. Zreikat is currently a full professor in the College of Engineering and Technology, American University of the Middle East, Kuwait (2015-present). He was previously a full professor in the College of Information Technology of Mutah University, Jordan (2004-2015). He obtained his PhD from the University of Bradford, UK, in 2003. His areas of research are the performance evaluation and resource management of wireless mobile networks and Internet of things (IoT) applications. He is a Senior Member of the IEEE, Computer Society member, member of the MOSEL research group, and member of the editorial advisory board for the book entitled Powering the Internet of Things with 5G Networks, published by IGI Global.