IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Digital Object Identifier 10.1109/ACCESS.2020.DOI

Early-stage Risk Prediction of Non-Communicable Disease using Machine Learning in Health CPS

RAHATARA FERDOUSI[1], M. ANWAR HOSSAIN[2], (Senior Member, IEEE), and ABDULMOTALEB EL SADDIK[3], (Fellow, IEEE)

[1,3]School of Electrical Engineering and Computer Science, University of Ottawa, Canada

[2]Department of Software Engineering, College of Computer and Information Sciences, King Saud University, Riyadh 11543, KSA

Corresponding author: M. Anwar Hossain (e-mail: mahossain@ksu.edu.sa)

ABSTRACT Cyber-Physical Systems (CPS) embed computation and communication capability into its core to regulate physical processes and seamlessly mediate between the cyber and the physical world for various control and monitoring tasks. Health CPS, a variant of CPS in the healthcare sector, acts as a health monitoring system to dynamically capture, process, and analyze health sensor data through integrated internet of things (IoT)-enabled cyber-physical processes. These systems can suitably support patients suffering from non-communicable diseases (NCDs) or who are at risk of suffering from those. Identifying the risk of NCDs, such as heart disease and diabetes, requires artificial intelligence (AI) techniques into the core of health CPS. Recently, there has been growing interest to incorporate machine learning into CPS, which can facilitate the disease classification, detection, monitoring, and prediction of several NCDs. However, there is a shortage of visible work that focus on early-stage risk prediction of these diseases. In this work, we propose a novel machine learning based health CPS framework that addresses the challenge of effectively processing the wearable IoT sensor data for early risk prediction of diabetes as an example of NCDs. In the experiment, a verified diabetic dataset has been used for training, while the testing has been performed on an artificially generated data collection from sensors. The experiment with several machine learning algorithms shows the effectiveness of the proposed approach in achieving the maximum precision from the Random Tree algorithm, which requires a minimum time of 0.01s to construct a model and obtains 94% accuracy to predict the probability of diabetes at an early point.

INDEX TERMS Cyber-physical Systems, Internet of Things, Machine Learning, Disease Risk Prediction, IoT Data Analysis

I. INTRODUCTION

Cyber-physical systems (CPS) promote the integration of IoT-enabled physical world with the computation-powered cyber world through seamless communication between them [1] [2] [3]. The interconnection between cyber and physical world are usually assisted by a feedback-loop control system [4] [3] [5], which enables CPS to be more adaptive to the changes in the physical world as sensed by various IoT sensing devices. The processing of diverse and dynamic data sources using different machine learning (ML) algorithms enable the CPS to exhibit higher intelligence [6] [7], which supports various application domains, including smart healthcare [8], transportation and others.

CPS has also emerged as medical CPS (MCPS) [9] [10] and smart health [11], to foresee the revolution in healthcare domain. These forms of CPS provide safety-critical functionality in patient-centric healthcare [12]. This is accomplished by collecting heterogeneous physiological parameters from patients and processing them with a goal to predict risk, detect abnormalities, or prevent from various non-communicable disease (NCD) conditions (e.g. coronary disease, diabetes, cancer, etc.). Processing of body-worn sensor data in a meaningful way using ML techniques has brought enormous complexity due to the diversity of health data [13], prompting new research in this domain [14] [7], [15].

The NCDs are not transferable and non-contagious [16], although more life-threatening than contagious diseases. According to the World Health Organization (WHO), NCDs are responsible for 71% of all deaths globally. However,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3094063, IEEE Access

the risk factors and determinants of these diseases, which are commonly known as epidemiological factors, are modifiable and controllable [17]. For example, obesity is an epidemiological factor that can cause NCDs like diabetes, stroke, hypertension, and kidney disease [18]. Therefore, the incidence of NCDs can be minimized by controlling these factors. Epidemiological factors of NCDs generally stem from physical inactivity, alcoholic habit, diets, and other conditions. Hence, pre-screening and preventive measures are the keys to respond to NCDs [19]. The value of health transformation, the empowerment of wearable sensors, and ML must be broadly acknowledged in the fight against losses due to NCDs [20] [21] [22]. However, few research efforts have been made in MCPS or HCPS domain to conduct early-stage risk prediction of NCDs, which is important to improve the health of the people by taking precautionary measures.

Among the many healthcare applications supported by MCPS and HCPS, patient monitoring from various perspectives has been the dominant one [12], [23], [24]. These include the work of remote patient observation [25] [26], activity monitoring [2], home health monitoring [27], heart health monitoring for cardiovascular disease [28], stroke detection [29], and epilepsy detection [30], to list a few. While such systems provide patient monitoring to a broader extent in a sensor-rich smart environment [25] [31], these are often used for disease classification and real-time alerting as a way of avoiding NCDs without any emphasis on early prediction of such diseases.

This paper reports our contribution in two-folds: first, we propose a closed-loop ML-powered HCPS for early-stage risk prediction of NCDs, considering diabetes as an example; and second, we incorporate the innovative concept of verified training dataset and dynamic test dataset, which have paved the way for applying ML on real-time data from wearable sensors. In order to support these contributions, we used different types of ML classification algorithms including Random Forest (RF), Decision Tree (DT), Naive Bayes (NB), Bayesian Net (BN), Multi-layer Perception (MLP), Support Vector Machine (SVM-Polykernel and SVM-RBFKernel), Logistic Regression (LR), Random Forest (RT), AdaBoost, Bagging, and K-th Nearest Neighbor (KNN). These algorithms have been suggested for diabetes prediction by researchers in existing work [32] [33] [34] [35]. In our proposed work, a particular real dataset of early-stage diabetes predictability is used for training purpose, whereas the testing is carried out using an externally supplied test dataset, which is dynamically generated from sensors in a simulated environment. We conducted several experiments that demonstrate that the proposed framework provides an effective mechanism for ML-based early-prediction of NCDs.

The rest of the work is organized as follows. Section II comments on the related work, while Section III provides the details of the proposed method. Section IV shows our experimental details along with a comparison and discussion on the results. Finally, Section V concludes the paper with a highlight to future work directions.

II. RELATED WORK

This section comments on existing work that are relevant to AI-based approach such as deep learning (DL), ML-based approach for smart health monitoring, AI-IoT convergence for healthcare, healthcare Internet of Things [36], CPS for smart healthcare, ML-based CPS, MCPS or HCPS for NCDs risk prediction, and more specifically ML in predicting diabetes risk with or without HCPS context [27], [37].

The authors of a recent survey [7] highlight the importance of incorporating intelligence into MCPS. The study reveals that the emerging health applications increasing need to include machine intelligence to provide innovative and smart services. The authors further describe the conversion of raw physiological inputs into functions and how those are used in ML, analyze the suitable ML algorithms, and describe how decisions are made and propagated to the user. In [23], the authors introduce a detail taxonomy for CPS in healthcare based on a comparative review of components and procedures. The taxonomy includes information about HCPS application, architecture, sensing approaches, data handling, computation, communication, security, and control, which can be consulted when developing HCPS applications.

The authors of a smart healthcare framework [11] highlight the importance of incorporating Gaussian mixture model-based classification for voice pathology detection that is used by physician for possible action. The findings of this study demonstrate how cloud and big data will improve the efficiency of healthcare system and provide smart healthcare solutions for the population. However, this work does not include any information of other disease prediction mechanism except for the voice pathology detection. The QoS issues have been studied in the context of remote healthcare in [26]. The work discusses the resolution of QoS challenges in urban healthcare big data system [38]. While it addresses the problems of healthcare and physical CPS systems, information about how IoT-sensor data can be analyzed intelligently for NCD predictions has not been made available.

The author in [24] proposes a CPS that incorporate localization information on the sensing, analyzing and sharing of patient data for continuous health monitoring, however, there is no indication of risk prediction of any particular disease in the work. In the area of general healthcare monitoring, the work in [25] shows a CPS implementation to monitor blood pressure (BP), blood glucose (BG), body temperature (BT), and heart beat rate (HR) based on embedded and cloud-based technology. This approach interconnects the communication, computation, and control aspect of CPS for continuous monitoring of patients and actuate remote treatment method when necessary.

The authors in [29] propose a CPS architecture for timely detection of stroke, a common NCD in patients, to minimize the risk in people. The system analyzes electroencephalography (EEG) data and connects to physician when it identifies stroke occurrence, and sends alerting message to the concerned personnel. However, it does not focus on early prediction of stroke.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3094063, IEEE Access

TABLE 1: Comparison of Related Work

Paper References	Application	Research Focus	Computational Model	Outcome	Comments
[7]	Healthcare monitoring	Incorporate AI into MCPS	–	Comparative study	AI incorporation into MCPS
[11]	Smart healthcare	healthcare data analytics	Data and architecture driven	Cloud-bigdata framework	Gaussian mixture model-based classifications
[26]	Remote healthcare monitoring	QoS modeling in CPS	Service oriented	Diagnoses suggestions	QoS for data collection, monitoring and decision-support
[24]	Patient monitoring	CPS Localization framework for voice and EEG signal acquisition	Data and architecture driven	Smartphone-based cloud-CPS framework	CPS localization for patient monitoring
[25]	Healthcare monitoring	Monitoring of BP, BG, BT, and HB	Embedded technology and cloud-assisted framework	CPS implementation of communication, computation, and control	Continuous monitoring, Remote actuation of treatment procedure
[29]	Stroke detection	Analyze EEG data in CPS	Data driven architecture	Stroke identification, Alerting Mechanism	Bridge gaps between CPS aspects
[27]	Home health monitoring	ML for EEG signal analysis	convolutional neural network (CNN)	Pathology identification	lightweight CNN model for health daignosis
[28]	Cardiovascular disease	CPS framework to connect ECG, PCG, and Lung Sound Sensors	Data and architecture driven	Model for data collection, processing and visualization	Mechanism to monitor cardiovascular patient and book appointment
[37]	Cognitive health	Application of 2D and 3D CNN	Extreme learning machine	Classification of emotional health	DL incorporating into cognitive health
[39]	Coronary heart disease	ECG data analysis to predict, monitor, and control the risk of heart disease	Data and cloud-driven framework	CPS-cloud with ANFIS neuro fuzzy inference system	Early risk prediction, connection to health professional, & suggest preventive action
[32], [33], [34], [35]	Diabetes diagnosis and prediction	Apply ML to analyze and predict diabetes	Data driven	Diabetes classification model	Diabetes prediction in non-CPS context
[40]	Connected health	Multimodal emotional health recognition	Gaussian mixture model and a support vector machine	Big-data framework for 5G	IoT for cognitive signal acquisition
Proposed work	NCDs prediction	ML-powered HCPS	Data and architecture driven	Framework for risk orediction of NCDs	Early-stage diabetes prediction in HCPS as an example

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3094063, IEEE Access

The processes of data collection, analysis, and visualization in CPS for cardiovascular disease have been demonstrated in [28]. The authors emphasize on constant tracking of patients' heart function using a smart phone and web-based interface. They enable CPS with cardiac signal processing capabilities based on ML and big data platform. Although the work does not elaborate on a specific prediction mechanism, it highlights the promise of machine and deep learning based CPS to support the identification and prediction of NCDs.

The work in [39] proposes a model to predict, monitor, and control the risk of coronary heart disease in CPS context. They authors use ANFIS fuzzy inference system to identify the different levels of risk assessment. They define 800 plus rules to determine the risk level and the consideration of additional attributes will require them to add even more rules, thereby increasing the overhead. Contrary to this work, our proposed method introduces verified training dataset against which ML classifier is built to predict the early risk level of diabetes from wearable sensor data.

As a summary of the above work, a comparative illustration is given in Table 1 that states key information from the reviewed papers in a convenient manner. It is evident from Table 1 that existing work focus on diverse aspects of smart healthcare. For example, the state-of-the art of CPS in smart healthcare, the architectural challenges of CPS, general health monitoring via HCPS, and the detection and monitoring of NCDs like stroke and cardiovascular disease. The table also includes the references of some general ML-based research [32] [33] [34] [35] that focus on the diagnosis and prediction of diabetes from different datasets, although not in HCPS context. We include those work to compare the results of diabetes prediction accuracy.

While HCPS for healthcare is currently in its early stage of adoption, there are limited experimental analysis that appears in existing work. A recent work in [39] does provide detail analysis for early prediction and monitoring of heart disease, more is needed to generalize a framework for early-prediction of other NCDs. The proposed work aims to contribute in this by providing an ML-enabled HCPS framework for the early-stage risk prediction of NCDs and demonstrates its objectives with the early prediction of diabetes.

III. PROPOSED METHOD

A. OVERVIEW

This research proposes an ML-powered HCPS system for the prediction of diabetes. Unlike traditional ML approach, which follows a longer training process associated with huge pre-processing, the proposed approach omits/ minimizes the pre-processing stage by introducing an epidemiological knowledge base. This includes the use of a verified training dataset approved by medical practitioner and rules to extract health data from raw sensor's data. For the testing phase, the proposal prescribes subsequent stages for obtaining a dynamic test dataset, which is produced from a combination of sensory and non-sensory data to fit the training data structure.

It should be noted that the involvement of medical practitioner into creating a knowledge base does not introduce delay per se in the training process. Rather, the verified training dataset make the end-user application robust and reliable, while providing a low-computational ML algorithms to process raw sensory data in IoT-embedded HCPS environment. This low-computational property of this approach is due to applying a verified training dataset to train the classifier and using a dynamic test dataset for evaluation. The detail of system processes relevant to training and testing phases appears in the following sections.

B. PROCESSES FOR TRAINING PHASE

1) Training Dataset Generation

The core of generating training dataset in HCPS is to define an epidemiology library [19] for disease risk factors from real patient data. The patient data can be collected from a direct pre-screening questionnaire or via other means, which are approved and overseen by the healthcare practitioners, who also verify the class level of data. The refined data are stored into a knowledge base. The practitioners-approved data have the potential to increase the level of acceptance for risk prediction of NCDs. To predict multiple NCDs through a single system, the epidemiology library using electronic healthcare records (EHR) can be constructed as a potential solution. The electronic healthcare records are used in many tele-healthcare systems.

2) Knowledge Base

A knowledge base includes verified datasets, ontology, and rules to label data [19]. For example, risk ontology, symptom and disease ontology, medical rules to determine attribute value, and other information which can be repeatedly used to serve data query. The use of an epidemiological knowledge base can accelerate the performance of the classification system. The proposed system uses the knowledge base in both the training and the testing phase. In the training phase, the verified dataset from the knowledge base has been used to train the classifier with several ML classification algorithms, while in testing phase the knowledge base has been used to assign rules and data labels and to extract features for predicting NCDs from sensor data.

3) Trained Classifier Building

The verified training dataset is used to train the classifier. Several popular ML algorithms [41] are used for NCD prediction. The outcome of the training phase is the trained classifiers. These classifiers are used to evaluate sensory data in the testing phase.

C. PROCESSES FOR TESTING PHASE

The proposed method innovates several processes for the HCPS testing phase. The goal is to generate a dynamic testing dataset from the raw sensor data and classify them for predicting diabetes. Similar process can be followed in

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3094063, IEEE Access

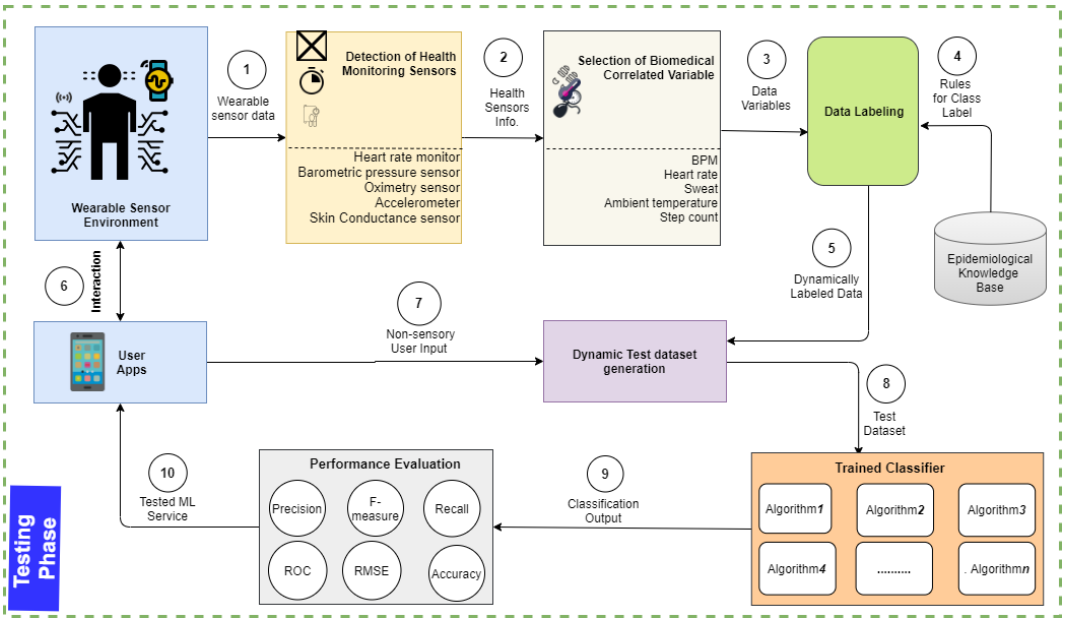

FIGURE 1: Proposed ML testing method in closed-loop health CPS environment

the case of predicting other NCDs, depending on the type of data system records and the classifiers it trains. The processes in this phase are depicted in Figure 1 and the data flow is marked with numbers to demonstrate a closed-loop model. The details of the processes are given as follows.

1) Detection of Health Monitoring Sensors

The wearable network includes sensors of diverse genres targeting different goals, such as health monitoring, disorder prediction, safety monitoring, home rehabilitation, activity monitoring, treatment assessment, and so on. Information about these sensors include sensor type, id, record type, manufacturer, and service information. These information can be extracted using Python command lines or dedicated tools. To detect whether the sensor is a health monitoring sensor or not, information about each wearable sensor in a specific network will be checked. If the model includes n wearable sensors then $W = w_1, w_2, w_3, ..., w_n$ is a list of sensors in an environment. For each wearable sensor w, an information list is extracted as $info\_list = i_1, i_2, .., i_m$. An instantiation of $info\_list$ can be seen in Figure 2.

2) Selection of Biomedical Correlated Data

The reading of the sensors will be used for selecting biomedical correlated variables. In particular, the correlation between sensor readings and the biomedical variable are assessed. Possible examples of biomedical correlated variables are beat per minute (BPM), sweat, step count, etc. To conduct this step, a pre-defined list containing the biomedical variables is checked. For example, if the list contains {bpm, step count, sweat, sleeping time} and the reading from a wearable sensor w is {bpm:60} then the variable bpm will be selected as a biomedical correlated variable.

{"sensor_name":"HeartRate","timestamp":"Sat Mar 12 00:15:51 PST 2020","sensor_data":{"bpm":66}}

{"sensor_name"."light", "timestamp".14.87545,"Thu Mar 10 17:28:46 EST 2020","sensor_data"."{less bright,17:00}}"

{"sensor_name":"battery","timestamp":"Sat Mar 12 00:15:51 PST 2020","sensor_data":{"charging":true}}

{"sensor_name":"step","timestamp":"Sun May 13 00:15:51 PST 2020","sensor_data":{"charging":true}}

FIGURE 2: Example of information list of sensors.

3) Data Labeling

The data labeling is performed by applying rules from the epidemiological knowledge base on the sensor data variables. The epidemiological knowledge base includes the features NCDs and data from general healthcare records (marked as 3 and 4 on Figure 1). These records provide the selection of epidemiological factors by the users through user interface at real time. For example, the knowledge base will include epidemiological factors (e.g. age, sudden weight-loss, and palpitation) with rules. These epidemiological factors are the filtered features. This kind of knowledge-driven feature selection is a low-computational approach to feature selection.

At this stage, the core contribution of this research has been introduced. The knowledge base including epidemiological factors of different NCDs will be applied to the sensor data variable. This will assign data labels dynamically to the regular sensor data. For example, if the weekly step count of a user is less than 2000, the value of obesity feature will be yes.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3094063, IEEE Access

Algorithm 1 Algorithm for Dynamic TESTING Data Generation.

class TESTING
    $feature\_name$
    $feature\_value$
end class
class RULE
    $match\_param\_name$
    $operator$
    $match\_param\_value$
    $decision\_param\_name$
    $decision\_param\_value$
end class
Initialize $TESTING$ list T;
Input list $health\_sensor\_info$, $RULE$ list $rule$;
for i=0 to $health\_sensor\_info.length()$ **do**
\ * Split sensor data which is in "data name: value" format.
\ * $find()$ is a Python function-returns index of a character in string.
\ * ':' is the delimiter, splitting value before and after ':' .
    $sensor\_data$ = $health\_sensor\_info[i].sensor\_data$;
    bio_cor_var = $sensor\_data[: sensor\_data.find(':')]$;
    bio_cor_val = $Int(sensor\_data[sensor\_data.find(':') + 1 :])$;
    for j=0 to $rule.length()$ **do**
        if $rule[j].match\_param\_name$ == bio_cor_var **then**
        \ *Check the sensor data with knowledge base rule.
            $Compare(bio\_cor\_val, rule[j].operator, rule[j].match\_param\_value)$
            if TRUE **then**
                $T[i].feature\_name$ = $rule[j].decision\_param\_name$;
                $T[i].feature\_value$ = $rule[j].decision\_param\_val$;
            end if
        end if
    end for
end for
Return T;

4) Dynamic Test Dataset Generation

The dynamic test data requires data from smart phone apps and of the labeled sensory data. The apps provide different information (e.g. age, gender) and other derived data such as genital thrush from the frequency of drinking time. The non-sensory data and dynamically labeled data along with the filtered features generate a dataset as the weekly records of a user. It should be noted here that non-sensory features such as age and gender will be used without any modification as it matches the feature format of verified training dataset.

The process of generating test data dynamically is given in Algorithm 1. The defined class TESTING has two attributes, $feature\_name$ to contain a feature name of the training dataset and $feature\_value$ to contain the value of a particular feature. The other defined class RULE has five attributes, which represent a rule in the knowledge base. An example of instantiation of an object of this class is: r=RULE($match\_param\_name$= 'bpm', $operator$ = $>$, $match\_param\_value$ = 80, $decision\_param\_name$ = 'Irritability', $decision\_param\_value$ = 1).

The algorithm takes health sensor information list as an input and initialize list T for output. In Python programming, a list is used as array. Therefore, we have used the term list and T as the list of TESTING instances. The contents of the information of sensors is presented in Figure 2. The sensor data of each sensor is then split into two parts with a delimiter. For example, sensor data = 'bpm:60' is split as bpm to be the biomedical correlated variable and 60 is its value.

Then the biomedical correlated variable is checked with the $match\_param\_name$ in the list of type RULE by the $Compare$ function. If the rule is satisfied then $decision\_param\_name$ is stored as the feature name of a TESTING instance and $decision\_param\_value$ is stored as the feature value. Finally the list T is returned as the dynamically created training data record.

5) Evaluation

Like any ML approach, the proposed model includes the necessary evaluation process. To evaluate the performance of the classifiers, we considered the widely accepted ML

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3094063, IEEE Access

TABLE 2: Training dataset details for the experiment.

Datasets	Total Instances	Positive Instances	Negative Instances	Source	Class Verified by
Training Set	520	314	186	Real data	Medical Practitioner

TABLE 3: Training dataset attributes

Attribute list	Attribute list (cont..)
1. Age	9. Visual blurring
2. Gender	10. Itching
3. Polyuria	11. Irritability
4. Polydipsia	12. Delayed healing
5. Sudden weight loss	13. Partial paresis
6. Weakness	14. Muscle stiffness
7. Polyphagia	15. Alopecia
8. Genital thrush	16. Obesity

evaluation measures against the dynamic test dataset. Finally, an evaluated version of ML services are provided to the end user application. The following parameters are used to measure the early-stage disease risk prediction. The results of evaluation with these parameters are provided in subsequent section.

- True Positive (TP) = NCD Risk identified correctly for those who are at risk.
- False Positive (FP) = NCD Risk free people identified incorrectly at risk.
- True Negative (TN) = NCD Risk free people identified correctly as risk free.
- False Negative (FN) = NCD Risk people are identified incorrectly as risk free.
- Correctly Classified = Percentage of instances classified correctly.
- Incorrectly Classified= Percentage of instances classified incorrectly.
- Kappa statistic (or kappa coefficient) = A kappa of 1 indicates perfect agreement, whereas a kappa of 0 indicates agreement equivalent to chance. The more the value close to 1 , the more better the performance.
- Root Mean Square Error (RMSE) = Standard deviation of the prediction errors. It measures how concentrated the data is around the line of best fit.
- TP Rate= TP/(TP + FN).
- FP Rate= FP/(FP + TN).
- Precision= TP/(TP + FP)
- Recall= TP /(TP + FN)
- F-measure=2*(Precision*Recall)/(Precision + Recall)
- ROC (Receiver Operating Characteristic) area = ROC curve is a plot of True positive rate‘and false positive rate. The closer the ROC area to 1.0, more accurate the classifier is.
- Accuracy=(TP + TN)/(TP + TN + FP + FN)

IV. EXPERIMENTAL ANALYSIS

A. EXPERIMENTAL SETUP

Following the method of this research, the experimental setup is divided into two phases. In the first phase, the training procedure is performed using a verified dataset of diabetes [42] [43]. This dataset has been collected with ethical approval and informed consent from real patients from a diabetic hospital. All data are collected from the patients prescription, where a medical officer identified a patient as diabetes potential. More specifically, patients who are recommended for clinical test are classified as positive. The use of this dataset is to predict the likelihood of diabetes at early-stage from common sign and symptoms such that potential loss of valuable life from diabetes can be minimized.

In the second phase, the test data has been produced from simulation and prototype to evaluate the performance of different classification algorithms. For the simulation network of wearable sensors, a sensor network was constructed by updating examples of cooja simulator using Contiki Operating System [44]. The sensors have been modified using Python programming language. Finally, the results have been compared with other existing works focusing on the context of diabetes risk prediction using ML.

B. EXPERIMENT FOR THE TRAINING PHASE

At this stage, a verified training dataset [42] is used as ground truth for the training purpose. There are total 17 attributes, including one class attribute. The detail information of the training dataset is provided in Table 2. The 16 attributes excluding the class attribute are taken from [43], which appear in Table 3.

The distribution of positive (clinical diabetes test prescribed) and negative (clinical diabetes test not prescribed) class in the training dataset is depicted in Figure 3. It can be observed that the dataset includes class variation for all the 16 attributes.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3094063, IEEE Access

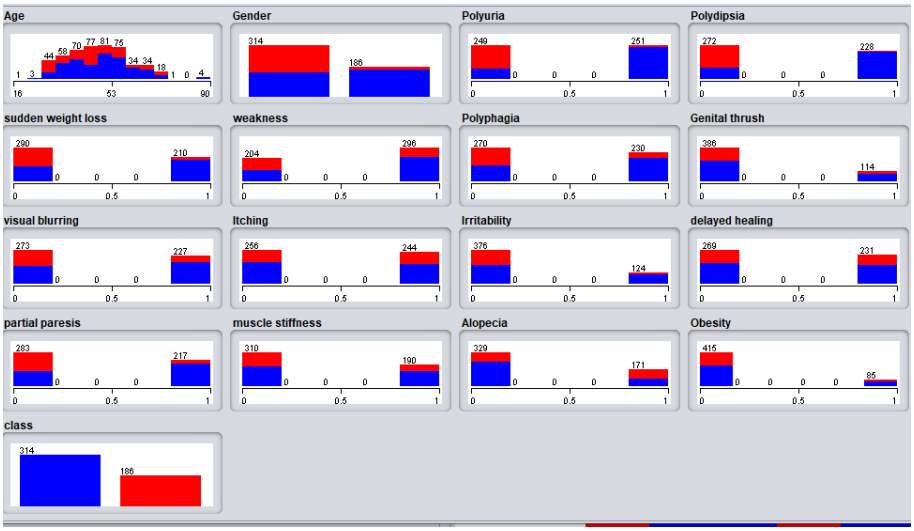

FIGURE 3: Attribute distribution in the training dataset.

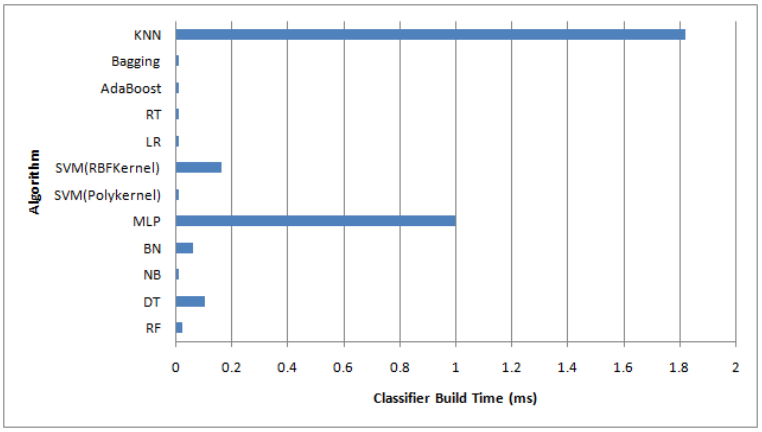

FIGURE 4: Time taken to build classifier with different algorithms during training.

Following the attribute distribution, the classifier was trained with training datasets applying 11 classification algorithms, which are RF, DT, NB, BN, MLP, SVM-Polykernel and SVM-RBFKernel, LR, RT, AdaBoost, Bagging, and KNN. The time to build model for each classification algorithm has been shown in Figure 4.

It can be observed from Figure 4 that the minimum time taken to build classifier is 0.01s for NB, SVM (PolyKernel), LR, RT, AdaBoost, and Bagging. The maximum time was required by KNN algorithm at 1.82s. Time to build model was recorded by using the whole training dataset. The classifier build time was expected to be minimal due to the use of verified training dataset, which can change in a different setup. However, the performance of the classification algorithms can not be summarized only from training time, a detail evaluation on the test dataset is therefore conducted and provided in the next section.

C. EXPERIMENT FOR THE TESTING PHASE

The testing phase includes several processes as explained before. Here we show the experimental aspects of the outcomes of these processes.

1) Identifying Health Monitoring Sensor from Sensor Payload

The information of the sensor are updated by the payloads. There are multiple wearable sensors in the network for different purposes. Unlike basic sensors like gyroscope or accelerometer, the modified sensors are able to provide more meaningful information. Table 4 provides detail information of the wearable sensors simulated for this work. The sensor_name represents the title of the sensors, which reflects the purpose of the sensors. The timestamps represent the time of capturing the sensor records. The sensor_data provides information about the reading of the sensors. Further elaboration of the sensors in Table 4 can be given as the following.

- The heart rate sensor provides bpm record.
- The eating sensor provides information about the number of food intake time.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3094063, IEEE Access

TABLE 4: Information list of sensors in the wearable sensor network.

Sensor ID	Information List
aaaa::212:7403:3:303	{"sensor_name":"Heartrate","timestamp":"Fri Mar 11 06:18:24 EST 2020","sensor_data":{"bpm":105}}
aaaa::212:740a:a:a0a	{"sensor_name":"Eating","timestamp":"Fri Mar 11 06:18:24 EST 2020","sensor_data":{"eat_count":10}}
aaaa::212:7402:2:202	{"sensor_name":"Light","timestamp":"Fri Mar 11 06:18:24 EST 2020","sensor_data":{"light_status":1}}
aaaa::212:7404:4:404	{"sensor_name":"Bluetooth","timestamp":"Fri Mar 11 06:18:24 EST 2020","sensor_data":{"btconnection":connected}}
aaaa::212:7405:5:505	{"sensor_name":"Stepcount","timestamp":"Fri Mar 11 06:18:24 EST 2020","sensor_data":{"step_count":2500}}
aaaa::212:7406:6:606	{"sensor_name":"Battery","timestamp":"Fri Mar 11 06:18:24 EST 2020","sensor_data":{"charging":70%}}
aaaa::212:7407:7:707	{"sensor_name":"Drinkwater","timestamp":"Fri Mar 11 06:18:24 EST 2020","sensor_data":{"drink_count":16}}
aaaa::212:7408:8:808	{"sensor_name":"Light","timestamp":"Fri Mar 11 06:18:24 EST 2020","sensor_data":{"light_status":1}}
aaaa::212:7409:9:909	{"sensor_name":"Skinrub","timestamp":"Fri Mar 11 06:18:24 EST 2020","sensor_data":{"skin_rub_count":35}}
aaaa::212:740b:b:b0b	{"sensor_name":"BP","timestamp":"Fri Mar 11 06:18:24 EST 2020","sensor_data":{"bp":90,50}}

TABLE 5: Detected health sensors.

Sensor ID	Sensor Name	Health Sensor? (yes=1, no=0)
aaaa::212:7403:3:303	{"sensor_name":"Heartrate"}	1
aaaa::212:740a:a:a0a	{"sensor_name":"Eating"}	1
aaaa::212:7402:2:202	{"sensor_name":"Light"}	0
aaaa::212:7404:4:404	{"sensor_name":"Bluetooth}	0
aaaa::212:7405:5:505	{"sensor_name":"Stepcount"}	1
aaaa::212:7406:6:606	{"sensor_name":"Battery"}	0
aaaa::212:7407:7:707	{"sensor_name":"Drinkwater"}	1
aaaa::212:7408:8:808	{"sensor_name":"Light"}	0
aaaa::212:7409:9:909	{"sensor_name":"Skinrub"}	1
aaaa::212:740b:b:b0b	{"sensor_name":"BP"}	1

TABLE 6: Biomedical correlated variable and sample rules from the epidemiological knowledge base to map sensor readings.

Sample sensor reading	Biomedical correlated variables	Rules	Decision
bpm:105	bpm	>100	Irritability=1
eat_count:10	eat_count	>5	Polyphagia=1
step_count:2500	step_count	<2000	Obesity=1
drink_count:16	drink_count	>15	Polydipsia=1
skin_rub_count:35	skin_rub_count	>30	Itching=1
bp:90,50	bp	<90,<60	Weakness=1

- The first light sensor is ON if the battery is charged and OFF otherwise.
- The Bluetooth sensors provide information about whether it is connected to the network or not. Another light sensor is on when the Bluetooth is connected.
- The step count sensor provides information about the number of steps completed by a person.
- Drink water sensors count the number of time water drunk by a person.
- The skin rub sensor provides information about the number of time the skin is rubbed by a person.
- The blood pressure (bp) sensor provides information about systolic and diastolic pressure (systolic, diastolic) in mm (Hg).
- In practice, smartwatch and fitness band has embedded sensors to provide more meaningful information [45]. These sensors are often made of basic sensors, which collect regular sensing data like pressure, temperature, movement, location, etc., and transform these data into more meaningful information like, drink water time, food intake, sleep time, etc.

The sensor_name is matched with a pre-defined list of health sensors and when a match is found the flag is set to 1 in the program to represent it as a health sensor. Based on the information in Table 4, it can be seen that there are six sensors found as health sensors in the wearable sensor network as shown in Table 5.

2) Obtaining Biomedical Correlated Variable from Sensor Payload

The sensor reading is matched with another list to identify the correlated variables. In Table 6, the identified variables from the sensor information can be seen, which are bpm, eat_count, step_count, drink_count, skin_rub_count and bp.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3094063, IEEE Access

TABLE 7: Test dataset details for the experiment.

Datasets	Total Instances	Positive Instances	Negative Instances	Source	Class Verified by
Test Set	216	81	134	Simulated data	N/A

TABLE 8: Filtered features from the epidemiological knowledge base and source of values.

Filtered attribute	Source of Value	Filtered attribute (cont..)	Source of Value
1. Age	Apps	9. Visual blurring	Apps
2. Gender	Apps	10. Itching	Sensor
3. Polyuria	Apps	11. Irritability	Sensor
4. Polydipsia	Sensor	12. Delayed healing	Apps
5. Sudden weight loss	Sensor	13. Partial paresis	Apps
6. Weakness	Sensor	14. Muscle stiffness	Apps
7. Polyphagia	Sensor	15. Alopecia	Apps
8. Genital thrush	Apps	16. Obesity	Sensor

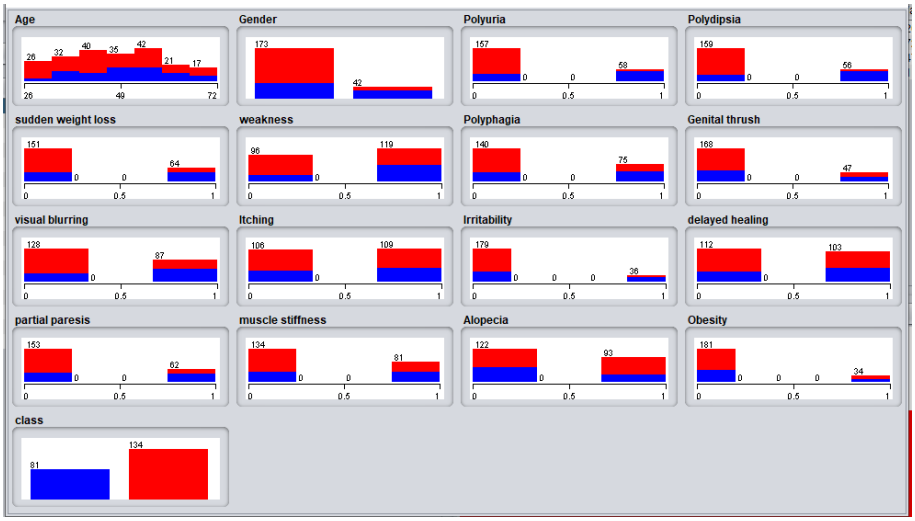

FIGURE 5: Attribute distribution in the test dataset.

3) Labeling Data

After extracting the variables from sensors, rule from the epidemiological knowledge base will be applied. The rules can be determined by domain experts or existing studies, such as in [46]. A sample of the rules for the biomedical correlated variables has been tabulated in Table 6. For our simulation, these rules have been used to label the data of different attributes.

4) Generating Dynamic Test Dataset

At this stage, the dynamic dataset for testing is produced using one week data from the sensor network, the detail of which is given in Table 7. More specifically, the features (attributes) set for predicting diabetes at an early stage in the diabetes dataset are selected as the filtered features. In the testing dataset, some of the data are obtained from the sensor network and some are from health applications. A health application was prototyped for collecting user responses like age, gender, and symptoms. In Table 8, the source of the field value for each filtered feature is shown.

As per the dynamically generated test dataset, the data distribution in the dynamic test sets is shown in Figure 5, which matches the attribute list of Figure 3 and shows the variation in terms of class distribution.

5) Evaluation

A thorough evaluation has been conducted on the test dataset, which is dynamically generated to evaluate the classification algorithms for early diabetes prediction. To evaluate the performance in details, several performance measures are considered as discussed in Section III-C5. First, we represent the correctly and incorrectly classified instances by each algorithm shown in Figure 6. It can be observed that RF, DT, MLP, RT, and KNN classified 91%-94% data correctly. The lowest correct classification is 81% by the NB and BN algorithms.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3094063, IEEE Access

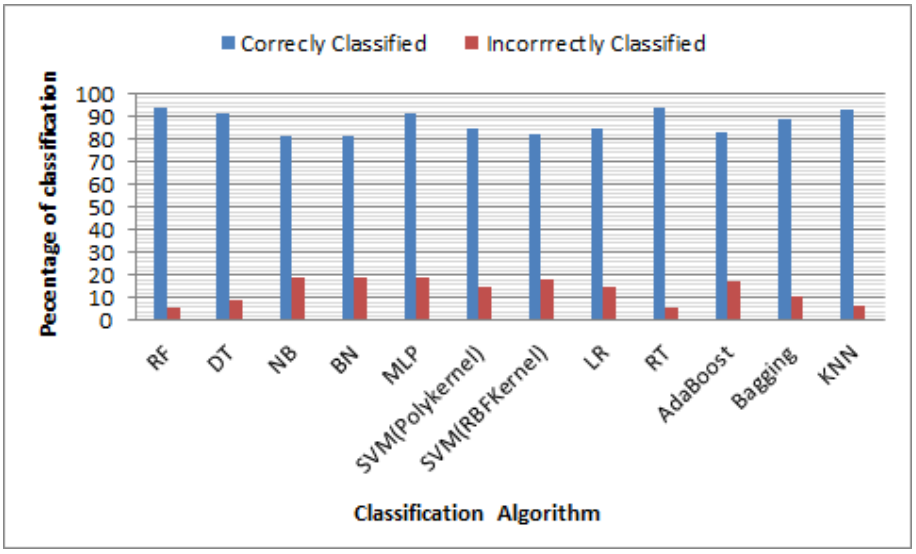

FIGURE 6: Correctly and incorrectly classified distribution in the test dataset.

TABLE 9: Confusion matrix for the classification algorithms.

(a) Confusion matrix of RF

Positive	Negative	Actual Class
69	12	Positive
0	134	Negative

(b) Confusion matrix of DT

Positive	Negative	Actual Class
64	17	Positive
3	131	Negative

(c) Confusion matrix of NB

Positive	Negative	Actual Class
58	23	Positive
17	117	Negative

(d) Confusion matrix of BN

Positive	Negative	Actual Class
54	27	Positive
13	121	Negative

(e) Confusion matrix of MLP

Positive	Negative	Actual Class
65	16	Positive
3	131	Negative

(f) Confusion matrix of SVM

Positive	Negative	Actual Class
63	20	Positive
13	21	Negative

(g) Confusion matrix of LR

Positive	Negative	Actual Class
62	19	Positive
15	119	Negative

(h) Confusion matrix of RT

Positive	Negative	Actual Class
66	15	Positive
0	134	Negative

(i) Confusion matrix of AdaBoost

Positive	Negative	Actual Class
60	21	Positive
17	117	Negative

(j) Confusion matrix of Bagging

Positive	Negative	Actual Class
63	18	Positive
4	130	Negative

(k) Confusion matrix of KNN

Positive	Negative	Actual Class
66	15	Positive
0	134	Negative

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3094063, IEEE Access

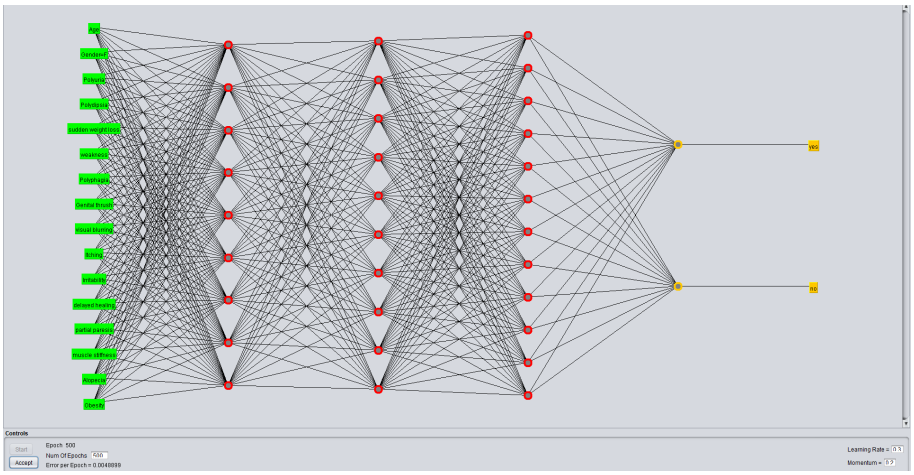

FIGURE 7: Neural network of MLP classification.

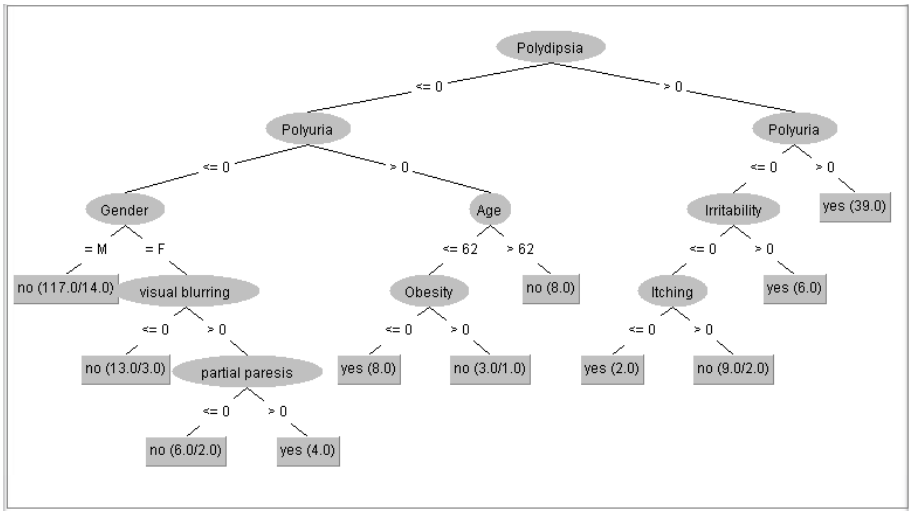

FIGURE 8: The tree from decision tree classification.

For more details, the confusion matrix for each algorithm has been given in Table 9. The confusion matrix represents predicted class and actual class of the prediction. The green color in the tables represent the predicted positive class for actual positive class, whereas the red color in the tables represent the predicted negative class for actual negative class. In this case RF leads again with 69 positive prediction among 81 actual positive class and 134 negative class among 134 negative class. That means no wrong prediction in case of negative class. RT also gives a decent outcome. The measures for NB is again low for this case accounting for 23 wrong prediction for the positive case. However, the MLP also provides a descent result by predicting actual classes correctly. The visualization of the MLP network has been demonstrated in Figure 7. While building this network in the WEKA tool we set the value of the hidden layers in the properties a, 9, 10, 12. Here, a is the default value calculated by a = (attributes + classes) / 2. In our case, a = (16 + 2) / 2 = 9. This setup provided very small error per epoch , which is 0.00049. The first layer in this figure contains the input and no computation is performed in this layer. The hidden layer includes computations to predict two classes.

The tree from the DT classification is depicted in Figure 8. The root of the tree is polydipsia, which then branches to polyuria and afterwards to reach the class attribute.

The Kappa statistics and RMSE comparison is provided in the Figure 9. These two statistical measures are considered widely for ML performance evaluation. The kappa statistics value of RF, RT, AdaBoost, and KNN is mostly closer to 1, which indicates the efficiency of these classification algorithms for this problem. On the other hand, the RMSE value of KNN, RT, MLP, and RF are the least, which proves the efficiency of those algorithms for the target prediction task.

To get a more detail view of the classifier performance, other accuracy measures like TP rate, FP rate, Precision, Recall, ROC area, and F-measure are illustrated in Figure 10. The highest value of these measures is approximately 0.93, 0.12, 0.94, 0.93, 0.93 and 0.93, respectively for multiple algorithms like RF and RT. The SVM performs worst among these algorithms. The SVM (RBFkernel) accounts for the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3094063, IEEE Access

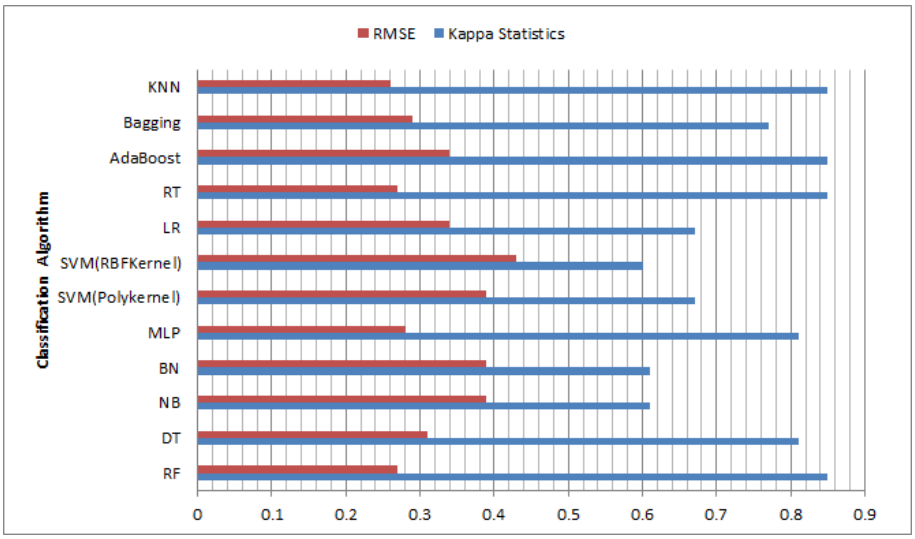

FIGURE 9: Statistical measures of classification algorithm.

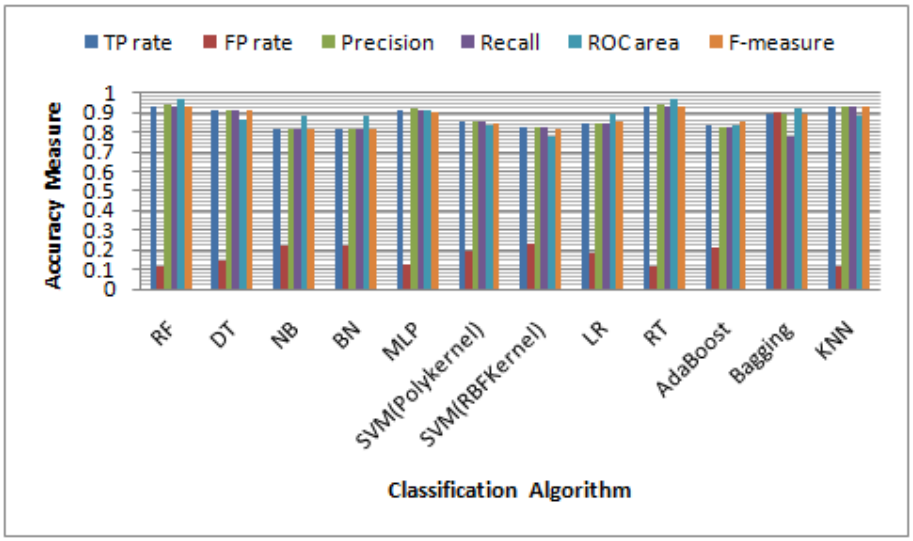

FIGURE 10: Accuracy measures of classification algorithm.

highest FP rate at 0.23 and the subsequent ones are for NB and BN at 0.22, and so on.

Based on all the analysis above, it is evident that RT performed best in terms of model build time during training and accuracy measures during testing. The RT take 0.01s time to build the classifier and exhibits 0.94 precision.

D. COMPARISON WITH EXISTING WORK

Existing work mostly consider clinical dataset for diabetic prediction, not for early-stage risk prediction of diabetes. Different work consider different datasets with diverse set of attributes. However, we took the context of diabetes and ML to compare our work that considers dataset and corresponding attirbutes for early-stage diabetes risk prediction. A comparison of the proposed work with the existing work has been outlined in Table 10.

The best accuracy for each algorithm is highlighted in Table 10. It is evident from the table that in most of the cases our proposed work provides the best accuracy. However, the work in [35] comes next providing best accuracy for three algorithms DT, NB, and SVM (Polykernal) and then in [34] for LR algorithm. Also, the - sign in the table cell represents that the corresponding algorithm is not used by the cited work. It can be observed from the table that each of existing work individually has used 3-4 classification techniques, whereas we analyzed our data with 11 classification techniques that have been popularly used for diabetes prediction in the literature. Therefore, comparatively the proposed work justifies the novelty in performance with respect to the referenced work.

E. SUMMARY OF RESULTS AND DISCUSSION

Overall, it is evident from the experimental results that although multiple algorithms built the model in minimum time

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3094063, IEEE Access

TABLE 10: Comparison of accuracy of our work with existing work.

Algorithms	RF	DT	NB	BN	MLP	SVM (Polykernel)	SVM (RBFKernel)	LR	RT	AdaBoost	Bagging	KNN
This work	94.02	91.02	81.03	81.02	92.3	85.14	82.28	84.67	94	83	89	93
[32]	-	-	73.82	N/A	71	-	-	69	-	-	-	83
[33]	76.3	73.82	-	-	N/A	N/A	65.1	-	-	-	-	-
[34]	N/A	70	-	75	77	74	-	98	-	-	-	-
[35]	N/A	93	92	N/A	91	91	82	N/A	-	-	-	-

within 0.01s, the accuracy of them varied significantly. For instance, the model is built in 0.01s by applying Bagging, AdaBoost, RT, LR, SVM(PolyKernel), and NB. However, the accuracy obtained by RF and RT is nearly 10% higher at 94% than SVM at 85.14%. Again, the RF provides the highest accuracy at 94.02% and ROC area at 0.97, but the time required by RF to build the model is two times more than RT. Though both RF and RT obtain the same value for TP rate, FP rate, Precision, Recall, F-measure and ROC accuracy measures, the less model built time supports RT to be the best algorithm for this experiment. This diversity of results provides interesting insights, such as a) Although several algorithms provide almost similar accuracy, the classification algorithms may require variable training time and b) For the prediction of NCDs, performance should be evaluated in both training and testing phase.

Overall, this research represents a new scope for early stage NCD prediction with modified wearable sensors. Interestingly, the epidemiological knowledge base made the approach sophisticated providing knowledge rules and NCD dataset. The dynamic labeling can solve the data labeling problem for classification in HCPS, one of the major problems in HCPS research field. Besides, this work represents the necessity for developing more advanced healthcare sensors, which can transform the health monitoring systems to complete healthcare systems. A NCD risk prediction closed-loop system can predict the risk of developing life-threatening disease (e.g. diabetes, thyroid, and stroke) at early stage and the public health of any community can be improved significantly, which can extend the life span of individuals as well.

V. CONCLUSION

In this work, the least unexplored field of healthcare, the early-stage risk prediction of NCDs through wearable technology in HCPS, has been studied. The use of a medical practitioner's verified training dataset in the framework has reduced the massive pre-processing stage of ML. In addition to this, a novel approach of dynamic test dataset generation from IoT sensors' raw data has been introduced. The multistage conversion of heterogeneous IoT sensor data into a meaningful dataset opens new door to predict the risk of NCDs from the low-level sensor data in HCPS. This has enabled the ML classification algorithms RF and RT to perform with 94% accuracy or more. Due to using perfectly refined training data, the classifier build time with training data becomes significantly low at 0.01s. Also, the comparison of the accuracy with other existing work demonstrates that the proposed framework performs the best for most of the classification algorithms considered. This work considers diabetes as an example of NCDs to demonstrate the novelty of the proposed mechanism. However, other NCDs such as stroke or thyroid can also be predicted with a proper epidemiological dataset, which can shape the further extension of this work.

.

REFERENCES

[1] E. A. Lee, "Cyber physical systems: Design challenges," in 2008 11th IEEE International Symposium on Object and Component-Oriented Real-Time Distributed Computing (ISORC). IEEE, 2008, pp. 363–369.
[2] F. Arafsha, F. Laamarti, and A. El Saddik, "Cyber-physical system framework for measurement and analysis of physical activities," Electronics, vol. 8, no. 2, p. 248, 2019.
[3] F. Hu, Y. Lu, A. V. Vasilakos, Q. Hao, R. Ma, Y. Patil, T. Zhang, J. Lu, X. Li, and N. N. Xiong, "Robust cyber–physical systems: Concept, models, and implementation," Future generation computer systems, vol. 56, pp. 449–475, 2016.
[4] K.-D. Kim and P. R. Kumar, "Cyber–physical systems: A perspective at the centennial," Proceedings of the IEEE, vol. 100, no. Special Centennial Issue, pp. 1287–1308, 2012.
[5] A. Darwish and A. E. Hassanien, "Cyber physical systems design, methodology, and integration: the current status and future outlook," Journal of Ambient Intelligence and Humanized Computing, vol. 9, no. 5, pp. 1541–1556, 2018.
[6] A. Håkansson, R. Hartung, and E. Moradian, "Reasoning strategies in smart cyber-physical systems," Procedia Computer Science, vol. 60, pp. 1575–1584, 2015.
[7] O. R. Shishvan, D.-S. Zois, and T. Soyata, "Incorporating artificial intelligence into medical cyber physical systems: A survey," in Connected Health in Smart Cities. Springer, 2020, pp. 153–178.
[8] S. U. Amin et al., "Cognitive smart healthcare for pathology detection and monitoring," IEEE Access, vol. 7, pp. 10 745–10 753, 2019.
[9] I. Lee, O. Sokolsky, S. Chen, J. Hatcliff, E. Jee, B. Kim, A. King, M. Mullen-Fortino, S. Park, A. Roederer et al., "Challenges and research directions in medical cyber–physical systems," Proceedings of the IEEE, vol. 100, no. 1, pp. 75–90, 2011.
[10] J. I. Jimenez, H. Jahankhani, and S. Kendzierskyj, "Health care in the cyberspace: Medical cyber-physical system and digital twin challenges," in Digital Twin Technologies and Smart Cities. Springer, 2020, pp. 79–92.
[11] M. S. Hossain, G. Muhammad, and A. Alamri, "Smart healthcare monitoring: a voice pathology detection paradigm for smart cities," Multimedia System, vol. 25, no. 5, p. 565–575, 2019.
[12] N. Dey, A. S. Ashour, F. Shi, S. J. Fong, and J. M. R. Tavares, "Medical cyber-physical systems: A survey," Journal of medical systems, vol. 42, no. 4, p. 74, 2018.
[13] S. U. Amin et al., "Deep learning for eeg motor imagery classification based on multi-layer cnns feature fusion," Future Generation Computer Systems, vol. 101, pp. 542–554, 2019.
[14] J. A. Stankovic, "Research directions for cyber physical systems in wireless and mobile healthcare," ACM Transactions on Cyber-Physical Systems, vol. 1, no. 1, pp. 1–12, 2016.
[15] A. Yassine et al., "IoT big data analytics for smart homes with fog and cloud computing," Future Generation Computer Systems, vol. 91, pp. 563–573, 2019.

[16] WHO. (2020, mar) Non communicable diseases. WHO. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases

[17] J. J. Miranda, S. Kinra, J. P. Casas, G. Davey Smith, and S. Ebrahim, "Non-communicable diseases in low-and middle-income countries: context, determinants and health policy," Tropical Medicine & International Health, vol. 13, no. 10, pp. 1225–1234, 2008.

[18] O. Simpson and S. G. Camorlinga, "A framework to study the emergence of non-communicable diseases," Procedia computer science, vol. 114, pp. 116–125, 2017.

[19] M. A. Hossain, R. Ferdousi, and M. F. Alhamid, "Knowledge-driven machine learning based framework for early-stage disease risk prediction in edge environment," Journal of Parallel and Distributed Computing, vol. 146, pp. 25–34, December 2020.

[20] M. Ambika and K. Latha, "Non-communicable diseases: an approach for prediction using machine learning technique," Int J Appl Eng Res, vol. 10, no. 55, pp. 806–810, 2015.

[21] E. Tambo, G. Madjou, and J. Ngogang, "Wearable sensors and healthcare informatics solutions in non-communicable diseases (ncds) prevention and management in africa," Journal of Health Medical Informatics, vol. 7, no. 01, pp. 1–6, 2016.

[22] S. Shinde and P. R. Rajeswari, "Intelligent health risk prediction systems using machine learning: a review," Int. J. Eng. Technol, vol. 7, p. 1019, 2018.

[23] S. A. Haque, S. M. Aziz, and M. Rahman, "Review of cyber-physical system in healthcare," international journal of distributed sensor networks, vol. 10, no. 4, p. 217415, 2014.

[24] M. S. Hossain, "Cloud-supported cyber–physical localization framework for patients monitoring," IEEE Systems Journal, vol. 11, no. 1, pp. 118–127, 2015.

[25] K. Monisha and M. R. Babu, "A novel framework for healthcare monitoring system through cyber-physical system," in Internet of Things and Personalized Healthcare Systems. Springer, 2019, pp. 21–36.

[26] T. Shah, A. Yavari, K. Mitra, S. Saguna, P. P. Jayaraman, F. Rabhi, and R. Ranjan, "Remote health care cyber-physical system: quality of service (qos) challenges and opportunities," IET Cyber-Physical Systems: Theory & Applications, vol. 1, no. 1, pp. 40–48, 2016.

[27] G. Muhammad, M. S. Hossain, and N. Kumar, "Eeg-based pathology detection for home health monitoring," IEEE Journal on Selected Areas in Communications, vol. 39, no. 2, pp. 603–610, 2021.

[28] D. Verma, "Cps-heart: cyber-physical systems for cardiovascular diseases," in Proceedings of the Workshop Program of the 19th International Conference on Distributed Computing and Networking, ACM. Varanasi, India: ACM, 2018, p. 26.

[29] A. Laghari, Z. A. Memon, S. Ullah, and I. Hussain, "Cyber physical system for stroke detection," IEEE Access, vol. 6, pp. 37 444–37 453, 2018.

[30] S. Mian Qaisar and A. Subasi, "Effective epileptic seizure detection based on the event-driven processing and machine learning for mobile healthcare," Journal of Ambient Intelligence and Humanized Computing, pp. 1–13, 2020.

[31] M. A. Hossain and D. T. Ahmed, "Virtual caregiver: an ambient-aware elderly monitoring system," IEEE Transactions on Information Technology in Biomedicine, vol. 16, no. 6, pp. 1024–1031, 2012.

[32] S. Selvakumar, K. S. Kannan, and S. GothaiNachiyar, "Prediction of diabetes diagnosis using classification based data mining techniques," International Journal of Statistics and Systems, vol. 12, no. 2, pp. 183–188, 2017.

[33] D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," Procedia computer science, vol. 132, pp. 1578–1585, 2018.

[34] A. K. Dwivedi, "Analysis of computational intelligence techniques for diabetes mellitus prediction," Neural Computing and Applications, vol. 30, no. 12, pp. 3837–3845, 2018.

[35] K. Sowjanya, A. Singhal, and C. Choudhary, "Mobdbtest: A machine learning based system for predicting diabetes risk using mobile devices," in 2015 IEEE International Advance Computing Conference (IACC). IEEE, 2015, pp. 397–402.

[36] L. Hu et al., "Software defined healthcare networks," IEEE Wireless Communications, vol. 22, no. 6, pp. 67–75, 2015.

[37] M. S. Hossain and G. Muhammad, "Emotion recognition using deep learning approach from audio-visual emotional big data," Information Fusion, vol. 49, pp. 69–78, 2019.

[38] M. Chen et al., "Urban healthcare big data system based on crowdsourced and cloud-based air quality indicators," IEEE Communications Magazine, vol. 56, no. 11, pp. 14–20, 2018.

[39] S. K. Sood and I. Mahajan, "A fog assisted cyber-physical framework for identifying and preventing coronary heart disease," Wireless Personal Communications, vol. 101, no. 1, pp. 143–165, 2018.

[40] M. S. Hossain and G. Muhammad, "Emotion-aware connected healthcare big data towards 5g," IEEE Internet of Things Journal, vol. 5, no. 4, pp. 2399–2406, 2018.

[41] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2016.

[42] "Uci machine learning repository," July 2020. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset

[43] M. F. Islam, R. Ferdousi, S. Rahman, and H. Y. Bushra, "Likelihood prediction of diabetes at early stage using data mining techniques," in Computer Vision and Machine Intelligence in Medical Image Analysis. Springer, 2020, pp. 113–125.

[44] B. Bagula and Z. Erasmus, "IoT emulation with cooja," in ICTP-IoT workshop, 2015.

[45] G. M. Weiss, "Wisdm smartphone and smartwatch activity and biometrics dataset," UCI Machine Learning Repository. [Online]. Available: https://archive.ics.uci.edu/ml/machine-learning-databases/00507/WISDM-dataset-description.pdf

[46] C. S. Koblenzer, "Itching and the atopic skin," Journal of allergy and clinical immunology, vol. 104, no. 3, pp. S109–S113, 1999.

RAHATARA FERDOUSI is a master's student in the School of electrical engineering and computer science (EECS) at the University of Ottawa. Previously, she has worked as a researcher of Advanced Systems and Software Research Lab (ASysLab), Khulna, Bangladesh. She also worked as a faculty at Metropolitan University, Sylhet, Bangladesh. She obtained her bachelor degree in Computer Science and Engineering Discipline from Khulna University, Bangladesh. Rahatara has been recognized as national and international innovator by a2i Programme, Prime Minister office in Bangladesh and UN Women. She has several awards for her innovative ideas to utilize technology for human welfare. As a new researcher she has authored/co-authored articles in reputed journals/conferences of IEEE and Springer.

M. ANWAR HOSSAIN (SM'17) is an Associate Professor in the Department of Software Engineering, College of Computer and Information Sciences at King Saud University (KSU), Riyadh, KSA. He obtained his master degree in Computer Science from the University of Ottawa, Canada, in 2005 and Ph.D. degree in Electrical and Computer Engineering from the same University in 2010. He obtained his bachelor degree in Computer Science and Engineering from Khulna University, Bangladesh. His current research includes Internet of Things, multimedia surveillance and privacy, Assisted Living, Artificial Intelligence, and Software Engineering. He has authored/co-authored over 100 research articles. Dr. Hossain has co-organized several IEEE/ACM workshops including IEEE ICME AAMS-PS 2011-13, IEEE ICME AMUSE 2014, ACM MM EMASC-2014, IEEE ISM CMAS-CITY2015, IEEE ICME MMCloudCity-2016, and IEEE ISM EMASC-2017 workshop. He served as a guest editor of Springer Multimedia Tools and Applications journal, International Journal of Distributed Sensor Networks, and Springer Multimedia Systems journal. He is an Associate Editor in several journals. He has secured several grants for research and innovation. He is a senior member of IEEE and ACM.

ABDULMOTALEB EL SADDIK (M'01–SM'04–F'09) is a Distinguished University Professor and University Research Chair in the School of electrical engineering and computer science at the University of Ottawa. His research focus is on the establishment of digital twins to facilitate the wellbeing of citizens using AI, IoT, AR/VR and 5G to allow people to interact in real time with one another as well as with their smart digital representations. He has coauthored 10 books and more than 550 publications and chaired more than 50 conferences and workshops. He has received research grants and contracts totaling more than $20 M. He has supervised more than 120 researchers and has received several international awards, for example, ACM Distinguished Scientist, Fellow of the Engineering Institute of Canada, Fellow of the Canadian Academy of Engineers and Fellow of IEEE, IEEE I&M Technical Achievement Award, IEEE Canada C.C. Gotlieb (Computer) Medal and the A.G.L. McNaughton Gold Medal for important contributions to the field of computer engineering and science.