


The Design and Implementation of a Semantic-Based Proactive System for Raw Sensor Data: A Case Study for Laboratory Environments

Mehmet Milli, Ozlem Varliklar, Musa Milli, Sanaz Lakestani


Abstract— In the last decade, raw sensor data from sensor-based systems, the area of use of which has increased considerably, pose a fundamentally new set of research challenges, including structuring, sharing, and management. Although many different academic studies have been conducted on the integration of sets of data emerging from different sensor-based systems until present, these studies have generally focused on the integration of data as syntax. Studies on the semantic integration of data are limited, and still, the area of the study mentioned have problems that await solutions. In this article; parameters (Carbon Dioxide (CO₂), Total Volatile Organic Compounds (TVOC), Carbon Monoxide (CO), Particulate Matter 2.5 (PM_{2.5}), Particulate Matter 10 (PM₁₀), Temperature, Humidity, Light), affecting laboratory analysis results and threatening the analyst's health, were measured in laboratory environments selected as “use cases”, and semantic-based information management framework was created for different sensor-based systems. Classical machine learning methods, and regression approaches which are frequently used for such sensor data, have been applied to the proposed sensor ontology and it has been measured that machine learning algorithm performs better on ontological sensor data. The most efficient algorithms in terms of accuracy and time were selected, and integrated into the proposed proactive approach, in order to take the selected laboratory environment's condition under control.

Index Terms— Sensor ontology, Semantic sensor web, Machine learning, Prediction on stream data, Supervised learning.


Mehmet Milli, is with Department of Computer Engineering University of Bolu Abant İzzet Baysal University, Bolu, Turkey, (e-mail: mehmetmilli@ibu.edu.tr).

 <https://orcid.org/0000-0002-0759-4433>


Ozlem Varliklar, is with Department of Computer Engineering University of Dokuz Eylül University, İzmir, Turkey, (e-mail: ozlem@cs.deu.edu.tr).

 <https://orcid.org/0000-0001-6415-0698>

Musa Milli, is with Department of Computer Engineering Turkish Naval Academy University of National Defense University, Istanbul, Turkey, (e-mail: musamilli@gmail.com).

 <https://orcid.org/0000-0001-8323-6366>

Sanaz Lakestani, is with Scientific Industrial and Technological Application and Research Center University of Bolu Abant İzzet Baysal University, Bolu, Turkey, (e-mail: sanazlakestani@ibu.edu.tr).

 <https://orcid.org/0000-0002-1661-7166>

Manuscript received Dec 12, 2022; accepted March 12, 2024.
DOI: [10.17694/bajece.1218009](https://doi.org/10.17694/bajece.1218009)

I. INTRODUCTION

ALTHOUGH SENSORS are defined differently in many studies, the most common definition is that it is known as devices that detect phenomena in the physical environment in which it is located [1]. In another definition, sensors are defined as devices that can convert chemical, physical, and biological values into digital values [2].

Sensors have evolved continuously since the day they emerged and reached such a capacity that it can be utilized in almost every application, presenting efficiency in size, cost and adequacy. As a result of all these developments, sensor-based systems have become the heart of many electronic systems today. The use of such systems in many areas has caused an exponential increase in raw data on the Internet. A demonstration of how the raw sensor data obtained from the sensor reaches consumers of data appears in Fig. 1.

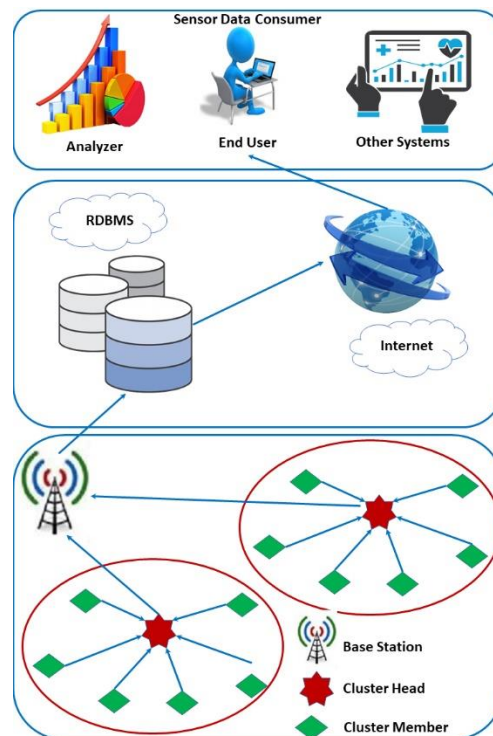


Fig. 1. The simple structure of a sensor-based system.

Most of the sensor data obtained from such systems on the Internet reach consumers without configuration. The unstructured presentation of sensor data causes a series of

problems that include sharing, interpreting, and managing data. Moreover, the sensor data is heterogeneous in nature because it bears different syntaxes, structures, and meanings in different systems [3]. The heterogeneity of the sensor data causes these data to remain application-specific, and hinders the management of independent sensor-based systems under a common infrastructure. An intermediate layer, independent of the application, enabling the sensor data semantically enriched to make it more useful is of crucial need.

In recent years, due to the reduction in size of sensors to a level suitable for use in every system, developments in the academic environment, and the continuous decrease in prices, sensor-based systems have rapidly spread to various aspects of daily life, particularly in industrial fields. The use of sensors in many fields has led to a significant increase in the raw data obtained from them. However, the lack of syntactic and semantic coherence among sensor data limits their sharing, reusability, and interpretability. The reusability, interpretability, and management of large-scale sensor data remain areas in need of effective solutions today. In the scope of this study, it is contemplated to address the mentioned issues by creating an ontology for raw sensor data.

Recently, researchers argue that semantic sensor web technologies can enrich the raw data obtained from sensors semantically and fill this intermediate layer [4-7]. Besides, a common framework is required for sensor-based information systems. Sensor data should be defined using Uniform Resource Identifiers (URIs) and delivered to sensor data consumers over HTTP [8]. In addition, sensor data should be encoded in formats that can be read by machines such as Resource Description Framework (RDF) and Web Ontology Language (OWL) so that they can be easily read and processed by machines. However, at this point, the lack of a comprehensive and understandable standard for the enrichment of sensor data around the world appears to be a major problem in the common manageability and operability of sensor systems.

The World Wide Web Consortium established the Semantic Sensor Network Incubator Group (SSN-XG) in 2011 to fill this intermediate layer and identified a set of standards for sensor data [9]. It has conducted many studies and defined certain standards for the semantic enrichment of raw sensor data obtained from SSN-XG sensor-based systems. The latest version of the Semantic Sensor Network (SSN), which is still used as a common framework in many studies today, was published in 2017 [10]. The core of SSN forms a lightweight but independent core ontology called SOSA (Sensor, Observation, Sample, and Actuator), which holds basic classes and properties. SOSA complies with the minimum interoperability limits, i.e. the sensor ontologies created with SOSA guarantees its sharing and interoperability with all other SSN and SOSA ontologies. Conceptual modules forming the infrastructure of sensor-based systems such as deployment, system, platform, procedure, and etc. are defined in the framework of SOSA and SSN. Some basic conceptual modules of SOSA/SSN are shown in Fig. 2.

The semantic sensor network is an application-independent framework that needs to be expanded with a certain concept and provides the manageability of the sensor systems on different platforms under a common infrastructure [11]. Shortly, SOSA/SSN is a model that allows the scope of the sensor ontology framework to be extended with other ontologies and concepts. For instance, in a biosensor application planned to be created in the field of medicine, a medical ontology, specific to the related field, including the technical medical terminology, classes, object properties and data properties can be employed to expand the ontological framework of SOSA-SSN.

A domain ontology that includes chemistry-related sensor measurements might import chemistry ontology, which includes chemical terminology (atomic number, orbital number, noble gas, element, etc.), classes, and object properties can be depicted as an example of the expansion of the SSN core ontology. The basic components of the SSN ontology are shown in Fig. 3.

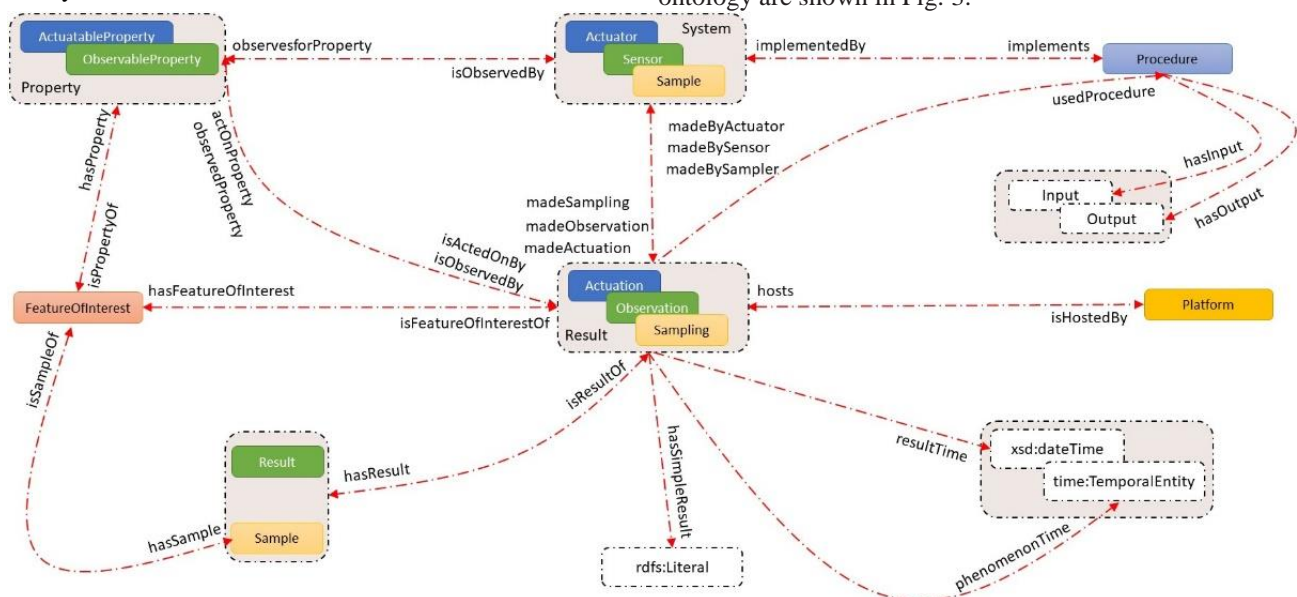


Fig. 2. Overview of the core structure of the SOSA classes, object properties, and data properties

The proposed ontology for laboratory environment parameters that affect the results of laboratory analysis and threaten the analyst's health during the analysis includes the general basic SOSA/SSN main classes. Only a few classes have been added to the basic SOSA/SSN framework. The added classes are described in detail in Section 3.2.

There is more than one purpose within the scope of the study. The main objectives of the study are listed below.

- Establishing a common infrastructure with a high capability to represent raw sensor data. Moreover, ensuring semantic integration of sensor data with each other by using ontological concepts such as Class, Object Property, Data Property. Hence, providing the capability to manage data obtained from different platforms, different systems, and different sensors under a common framework.
- To establish a system that provides real-time monitoring and control of laboratory environment parameters that negatively affect the laboratory analysis results and threaten the analyst's health.
- Determining the best algorithm for the designated laboratory environment parameters by using classical machine learning algorithms on ontological sensor data. And accordingly, detection of unforeseen environmental situations thanks to the ontological based proactive system created, and avoiding unwanted situations by executing appropriate action plans in time.

In this study, it is considered that creating an ontology of sensor data will contribute to the literature. These contributions roughly include: (i) Establishing a common framework for inherently heterogeneous sensor data, (ii) facilitating the shareability and reusability of sensor data across different platforms, hence enhancing the sustainability of sensor-based systems, (iii) ensuring machine readability of sensor data by encoding it in structural languages such as RDF and OWL, (iv) enriching sensor data semantically to make it machine-interpretable, (v) finally, in this study, an example of ontological sensor data was created, and commonly used regression models and machine learning algorithms were tested to demonstrate which ones can be applied to ontological sensor data in the literature.

The remainder of the article is organized as follows. In Section 2, previous studies in the field of sensor ontology are examined, and the differences between those and the current ongoing study are clearly revealed. Setting up systems infrastructure, creating sensor nodes, and use case are presented in Section 3. The data collection, the experiments to prepare data for the machine learning algorithm, and choosing the appropriate machine learning algorithms for the proposed sensor ontology are presented in Section 4. Section 5 describes the comparison of machine learning algorithms, determination of the most suitable algorithm in every aspect, and integration into the proposed proactive system. Finally, the results and future studies are discussed in detail in Section 6.

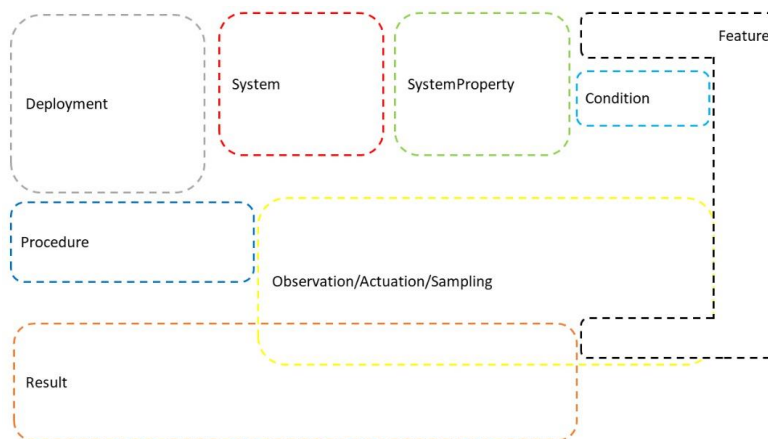


Fig. 3. Basic conceptual ontology modules of SOSA/SSN frameworks [9].

II. RELATED WORK

The concept of sensor data ontology was first introduced by Avancha et al. [12]. Since 2004, many studies have been carried out in this field, and sensor ontology has become an area of study that attracts more attention. Considering the components (machine learning, semantic web technologies, wireless sensor networks) that form the basis of the proposed study, there are many studies in the literature. Therefore, it is possible to classify the literature review under 3 headings by selecting articles that are similar to this study.

The works in the first group focus on the integration of machine learning algorithms built on data from wireless sensor networks (WSN). In this category, studies focusing on machine learning algorithms processing sensor data and excluding semantic enrichment approaches are argued. In this

context, many studies have been administered in different domains in the last 20 years. These studies cover the applications of machine learning approaches in the field of health in [13-15]. In [16, 17] there are studies in which machine learning approaches are applied in the field of environment and agriculture. In addition to these, machine learning approaches have been used in areas such as smart cities [18], security [19 -21] where WSN's are frequently utilized. Studies in this area are not assessed in detail, as they are a bit far from the proposed study. The major difference between these studies in the first group and the proposed study is that the sensor data collected is not enriched by using the ontological concepts. The best advantage of the proposed system is that it enables the management of ontologically oriented application-specific sensor-based systems before the emergence of the SOSA/SSN common framework.

In the second group, studies focusing on structuring sensor data to be managed under a common framework are considered. Although the semantic sensor web is beneficial in ensuring analytical integration between different sets raw data, the complexity of semantic techniques is often unacceptable for some end-users and data consumers due to the long processing time. The suggested system in [22] proposes IoT-Lite to reduce complexity and shorten transaction times. The IoT-Lite contains a simple example of semantic sensor ontology. The greatest feature of this sensor ontology is an approach that provides interoperability of sensor data on heterogeneous Internet of Things (IoT) platforms and includes minimum concepts and relationships that can respond to most end-user questions in a reasonable time. In the work mentioned in [23], a semantic sensor network has been used to solve interoperability problems of different platforms and devices in an e-health system. Apart from these, Kuster et al. [24], Wang et al. [25], Ali et al. [26] proposed different semantically based architectures to describe sensor information collected from different environments.

In these studies, the focus is on the management of sensor data feeding from different systems under a common infrastructure. The major difference between these studies and the proposed study is that machine learning algorithms are not operated on the sensor data of which ontology is created. In other words, these systems only perform real-time monitoring in real-world applications. In the suggested system, one of the main objectives is to find the most suitable machine learning approach for the proposed ontological sensor system.

In the third group, the studies cover the application of machine learning approaches to semantic sensor data. The studies closest to the proposed study are examined in this group. The system proposed in [27] mentions a sensor ontology which is presented using the World Wide Web Consortium's (W3C) SSN frame. Adeleke et al. developed a statistical machine learning-based prediction model using this proposed sensor ontology. In the respected study, in order to predict an unhealthy situation in the near future, their models are evaluated on PM_{2.5} and PM₁₀ values. 5 different classification algorithms are applied to ontological sensor data in their studies. By comparing these algorithms, they claim that the most effective algorithm on PM values is the Multilayer Perceptron.

In the work mentioned in Onal et al. [28], another semantic sensor web-based proactive system is presented. This system has been applied and evaluated for clustering and sensor anomaly detection using a public data set. In this study, the LinkedSensorData and LinkedObservationData dataset containing different weather parameters such as air temperature, wind speed, relative humidity, pressure, and visibility are used. LinkedSensorData is an RDF dataset that describes approximately 8000 air sensor information. The K-means algorithm, which is widely used for proactive systems in the literature, has been chosen as the appropriate model in this system.

The studies that are the most similar to the proposed study in terms of technology and scope are evaluated in this group. Studies under this category have also created a semantic-based

framework for the definition of sensor information, and classical machine learning approaches have been performed on ontological sensor information. The main purpose of SSN is to create a common identification frame for sensor information from different platforms, different domains, and different sensors. However, in these studies, the number of platforms, sensors, and domains are limited and the capacity of SSN to represent sensor information in different systems, platforms, and domains could not be fully utilized. In the proposed study, 3 different environments, 4 different platforms, 5 different sensors are used and 8 different parameter values are measured. In previous studies, machine learning algorithms applied to ontological sensor data are limited in number, so in this study, the number of algorithms running on sensor data is increased. Another difference is that many studies focused on either regression or binary classification. In this study, regression and binary classification approaches are evaluated together.

Apart from all these studies explained above, the study field of semantic sensor data has been expanded to include increasing scalability, aligning ontologies, and integrating them into the Internet of Things platforms. Al-Baltah et al. [29] have focused on the semantic integration of heterogeneous sensor data from different systems and sources between machines. One of the biggest hurdles in integrating heterogeneous sensor data is scalability. Therefore, the researchers propose a scalable semantic data aggregation framework that aims to improve the scalability of data integration in their models and to detect and reconcile unit of measure conflicts. In this study, to prove the feasibility of the proposed framework, real sensor data was collected and carried out as a web application. Experimental results on the use-case conducted by the researchers show that their proposed framework improves the scalability of data aggregation between heterogeneous sensors data. Another result of the proposed framework is that it is effective in detecting and resolving measurement unit conflicts.

Another area where semantic sensor technologies have been used recently is the IoT. IoT sensors continuously generate large volumes of observed stream data. Processing this data and integrating it into other systems may sometimes require going beyond classical approaches. For this reason, recently, many researchers argue that integrating semantic web technologies into IoT systems plays an active role in the instant decision-making mechanisms of the proposed studies [30-32]. These researches focus on real-time processing and interpretation of sensor flow data by integrating different semantic descriptions into the proposed frameworks. It is thought that efforts to integrate Semantic Web technologies into IoT frameworks will increase day by day due to their data integration capabilities.

The subject under study is actually closely related to smart buildings as well. Monitoring indoor air quality and ensuring actions are taken in inappropriate situations are subtopics of creating smart buildings. The success of smart building designs relies on bringing together expertise from various fields. Coordinating processes that require different fields of expertise involves integrating data and concepts obtained from

these fields in an appropriate manner. The study conducted in [33] includes a taxonomy on semantic web technologies and categorizes ontology studies for smart buildings into three main categories and several subcategories. Indoor air quality measurement with sensor data and intervention when necessary has been classified among dynamic features. Because indoor air is a spatiotemporal feature that varies over time and space. In the study, in addition to the developed taxonomy, a new ontology that integrates the static and dynamic features of smart buildings has been proposed. Our proposed study can be described as a more comprehensive and practical version of the dynamic part of the ontology created in the mentioned study.

Ontologies, which are useful tools for integrating data collected from different fields, also appear in smart city models, just like in smart buildings. Smart city concepts have been researched in order to alleviate traffic difficulties and the associated problems [34]. In the "Understanding Traffic Flows to Improve Air Quality" (TRAF AIR) project proposed by Desimony et.al, cameras recording the traffic situation and sensors measuring environmental parameters have been placed on the roadsides in Modena (Italy) and Zaragoza (Spain). The data collected through cameras and sensors are processed and stored in the TRAF AIR database in CSV format. Semantic relationships between data in CSV format have been established using appropriate ontologies. The aim of the project is to examine the impact of traffic flow data on air pollution in cities and to predict future air pollution. The main difference between the study conducted within the TRAF AIR project and our study is that, in the TRAF AIR project, environmental parameters are collected in an open-air environment, while in our current study, environmental parameters are collected in an indoor environment.

Poor indoor air quality has long been a major concern for human health. Recently, especially with the SARS-CoV-2 pandemic, studies have been conducted to control indoor air quality and improve indoor air quality with appropriate actions when necessary [35, 36]. During the SARS-CoV-2 pandemic, 24/7 lockdown has been enforced for 45 days in Spain. Domínguez-Amarillo et.al [35] conducted indoor air parameter measurements (CO₂, PM_{2.5}, NO₂, TVOC) in four different types of homes in Madrid before and during the lockdown. The study reveals that, during the lockdown, while outdoor air quality improved, indoor air quality deteriorated dramatically. In the study, measures to be taken during a full or partial lockdown, especially for individuals with respiratory problems, have been evaluated. In our proposed work, measurements of a greater number of environmental parameters have been conducted. In our study, a sensor ontology concept was proposed to facilitate the integration of the obtained data. Additionally, using various machine learning algorithms, the environmental air quality was assessed to determine whether it falls within normal values for human health.

Unlike outdoor air, due to limited circulation indoors, air pollutants tend to accumulate continuously in the environment,

leading to a faster penetration of disease-causing organisms. Monitoring, controlling, and taking necessary actions for air quality have become even more crucial, especially with the SARS-CoV-2 pandemic. In this context, Mumtaz et.al [36] measured indoor air quality using gas and particle sensors, and they established an indoor air quality monitoring system. The study involves measuring air quality, generating alerts if any measured parameter exceeds a threshold value, and predicting future air quality. As a result, the study contains several preventative strategies. Similar to our work, a sensor node measuring 8 different parameters affecting air quality has been created. The main difference in our study compared to this study is the use of a conceptual sensor ontology for the integration of data.

III. MATERIALS AND METHODS

A. Sensor Nodes

In order to measure the values of parameters determined in the selected use case, 4 different nodes to perform 4 different tasks have been established. These sensor nodes are named Type A, Type B, Type C, and Type D and the purpose of installation and fundamentals components are given below. Arduino Uno is used as a microprocessor in all sensor nodes due to its ease of use and low cost. Considering transmission distance, energy consumption, and compatibility with Arduino Uno, the nrf24l01+ antenna is chosen as the communication device. In order to reduce the load on the nodes and to provide the flexibility of deployment during the distribution of the sensors in the environment, two different sensor nodes are installed, and the sensors are placed on them.

Type A Sensor Node (Gateway Node): The gateway node is the most important node in the network, as it is the one to collect the data and transmit to the base station. In cases where the Type A sensor node fails to function due to physical obstacles or any problem arising from its electronics, or if communication with other nodes is interrupted, all data communication in the network stops. Thus, the Type A sensor node is vital for the system. No sensor was placed on it as no measurement in the environment is expected from it.

Type B Sensor Node (Sensor Node 1): In the proposed project, 5 different sensors are used to measure 8 parameters. These sensors are integrated into the two nodes, measuring an equal number of parameters. The digital humidity and temperature Sensor (DHT22), which measures the temperature and humidity parameters in the environment, and the combined CO₂ and TVOC sensor (CCS811) that measures the carbon dioxide and total volatile organic compounds, are integrated on the Type B Sensor Node.

Type C Sensor Node (Sensor Node 2): Another sensor node that makes measurements in the environment is the Type C sensor node. MQ-7 sensor measuring carbon monoxide, Nova SDS011 sensor measuring PM_{2.5}, and PM₁₀ values, and light-dependent resistance (LDR) sensor measuring light intensity in the environment are integrated into this node.

Type D Sensor Node (Repeater Node): After the nodes are installed in the measurement environment and WSN is established, a communication problem occurs due to the distance and obstacles between some nodes. In order to solve this communication problem and to ensure healthy data communication, repeater nodes are placed to strengthen the received signal and to enable the data received from the node to reach the gateway node. The sensors used, the nodes created, the technical infrastructure of this network, the characteristics, and detailed description of this system used are available in the previous study of the research team [37].

B. Sensor Ontology

The SOSA/SSN provides an application-independent common framework that needs to be expanded with specific concepts and opportunity to manage sensor data for different domains. The concepts to be added can be classes, object properties, data property, or individuals depending on the application. In the proposed project, the core SSN ontology for the ontology of laboratory environment parameters is expanded by adding some classes, object properties, and individuals. This ontology is designed with the Protege [38] ontology editor developed by Stanford University. Protege is a free open source framework that provides an interface for users to review ontologies. The Protege 5.5.0 editor has the capability to create classes and subclasses, define and visualize the relationship between classes to extend the SSN ontology.

Since this article focuses more on seeking the most appropriate machine learning approaches on ontological sensor data for proactive system design, the creation of sensor ontology is not explained in detail. Technical information on the proposed sensor ontology is available in the article previously written by the project team [37]. SSN core sensor ontology has been

expanded to represent the environmental parameters that affect the analysis performed in the laboratory environment used and the indoor environment parameters that affects the health of the analyst. This extension includes appropriate classes, object properties, data properties, and instances. The following example is given in order to better understand the proposed sensor ontology. In the core SSN ontology, the most significant concept is the "sosa:Observation" class, as it represents the sensor value and measurement date and time with the data properties attached to it. Fig. 4 below shows an example of an extended sensor ontology from the point of view of the "sosa:Observation" class in the proposed sensor ontology.

"sosa:Observation" is the indicator representing the value of the property of a "sosa:FeatureOfInterest", or computing through a "sosa:Procedure". The algorithm connects to "sosa:Sensor", subclass of "ssn: System" class with "sosa: madeBySensor" object property, to understand what shapes "sosa:Observation". In the above illustration, Nova SDS011 sensor used in the project is given as an example. An individual of the class "sosa:Observation" measured by this sensor is shown in Fig. 4. Each measurement is given a unique value consisting of 32 characters and represented by it. So that, data consumers can access each sensor value they want to display with this unique id. The PM_{2.5} value measured by the Nova SDS011 value is "xsd:double" 7.73. As illustrated, the measurement date and time are "xsd:dateTime" 2019-08-30T06: 00: 00 + 03: 00. Since there is a good number of units for the same or different parameters in the literature, the "MeasurementUnit" class has been added to the basic SOSA/SSN framework to avoid unit complexity. The unit of the value measured by the Nova SDS011 sensor given in the example is assigned as "PartsPerMillion", which is frequently used in the literature.

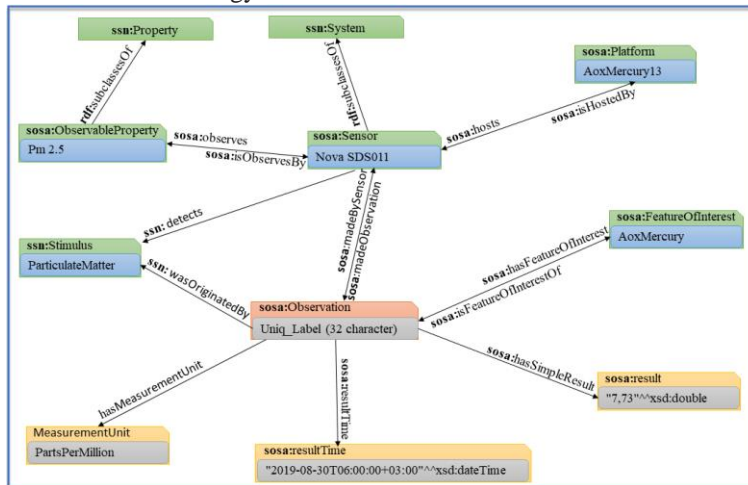


Fig. 4. Overview of some SOSA/SSN classes and properties from the sosa:Observation class perspective.

Looking at the other concepts in the given example, to explain which parameter is measured by "sosa:Sensor" class, a link is given to "sosa:ObservableProperty", which is a sub-class of "sosa:Property" class, with "sosa:observes" object property. In this example, it is seen that the parameter measured by Nova SDS011 sensor, which is a member of the "sosa: Sensor" class, is PM_{2.5} "sosa:isFeatureOfInterestOf" object property is given as a link to "sosa:FeatureofInterest" classes to explain to

which environment the value "sosa:Observation" is associated. "sosa:FeatureOfInterest" class is the area or environment where you want to measure. In this study, 3 laboratories that are frequently used in Scientific Industrial and Technological Application and Research Center (SITARC) have been selected as the measurement area. One of them is the AoxMercury laboratory where various analyses are carried out. To summarize the example given above, in the

AoxMercury measurement area, the value measured by the Nova SDS011 sensor on the AoxMercury13 platform at 06:00 a.m. on 30.08.2019 is 7.73 ppm.

C. Use Case

The proposed sensor ontology is created using the sensor data collected in the SITARC within the Bolu Abant Izzet Baysal University (BAIBU). Data collection has been carried out in 3 laboratories frequently used in SITARC. These laboratories are MaldiTof, AoxMercury, and Chromatography laboratories. In these laboratories selected as Use Cases, microorganism identification, proteomic analysis, bacteria count, fatty acid analysis, anion-cation determination, total halogen determination, solid-phase extraction, etc. analyses are done frequently.

During analyses, both the environmental parameters that will affect the analyst's health and the environmental parameters that will affect the analysis result must be monitored instantaneously in order to be kept under control. According to the report of the World Health Organization (WHO) [32] one of the most important causes of disease and death in the world is an unhealthy living environment. Therefore, avoiding unhealthy conditions and monitoring the working environment effectively to keep the environmental parameters under control emerges as a serious issue.

In this study, a total of 8 parameters: temperature, humidity, CO₂, TVOC, CO, PM_{2.5}, PM₁₀, and light intensity are measured by 5 sensors. For this, a total of 8 sensor nodes, including 1 Type A, 3 Type B, 3 Type C, and 1 Type D nodes, are established and deployed to measurement environments. 1 Type B and 1 Type C sensor nodes are placed in every 3 laboratories selected as measurement areas, one for each sensor. Type A sensor node (Gateway) is placed in AoxMercury Laboratory because it is close to the midpoint of all nodes. Once the sensor network is established, it is realized that there are communication problems between the Gateway and Type C sensor node, due to distance and physical obstacles such as walls, tables, and devices in the Chromatography laboratory from time to time. This problem is solved by placing a Type D sensor node between these nodes and strengthening the signal.

IV. EXPERIMENTS

A. Collecting Data

After placing the sensor nodes in the measurement area and sending the data properly, the data collection process is started on 29.08.2019 at 16:05. Each sensor in the installed system is programmed to measure an average per minute and send it to the gateway. The hourly average of the collected data is added to Apache Jena Fuseki, which is frequently used as a triple database (Apache Software Foundation) [39]. Jena Fuseki is a SPARQL Protocol and RDF Query Language (SPARQL) server. In addition, it has been preferred as a triple database in this project as it provides a clear user interface for server monitoring and management.

The data collection process has been terminated on 12.10.2019 due to the annual maintenance of the devices in the laboratory. A total of 45 days of uninterrupted data has been collected at the selected measurement sites. Between these dates, each sensor has made approximately 62,000 measurements, and a total of approximately 1,500,000 measurements have been made. Theoretical and practical training have been given twice in the first 10 days of September and October in the laboratories specified between the dates of measurement, and the 3 laboratories where the measurement is made have been used. This situation has been beneficial for the project results in terms of observing what kind of changes may occur in the parameters during the analysis and training in the laboratory. Daily average values of temperature and CO₂ between the measurement dates are shown in Fig. 5 and Fig. 6 respectively.

The graph in Fig. 5 is given as a box-whisker plot to clearly show the central position and spread of the mean of temperature data collected. Although the low values of some parameters such as temperature during analysis have a positive effect on analysis studies, it negatively affects the health of the personnel, especially in long-term analyses. Especially in MaldiTof and Chromatography laboratories, the ambient temperature must be below 18 °C for a proper analysis activity to be carried out. However, considering the health of the personnel, it is important to keep the temperature in these laboratories within a narrow range. Although there are air conditioners, keeping the ambient temperature at appropriate levels that do not expose a threat on human health and not negatively affect the analysis results in laboratories, is more complicated than in other environments.

The graphic in Fig. 6 shows the daily average CO₂ level in the laboratories selected for the measurement area within the specified date range. Especially during the dates of theoretical and practical training, it is seen that the amount of CO₂ in the environment exceeded the value of 1000 ppm determined by the WHO health organization as a reference value for indoor environments. It has been observed that the value of many parameters measured within the scope of this study increased during the dates of formal education. The reason for this increase is thought to be directly related to the amount of gas released as a result of the analysis performed in the experiments and increasing the human activity in the environment.

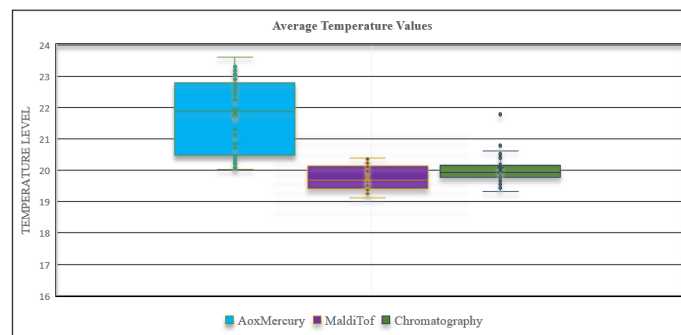


Fig. 5. Box and Dispersion (spread) graph of average temperature values between 29.08.2019 and 12.10.2019 in laboratories.

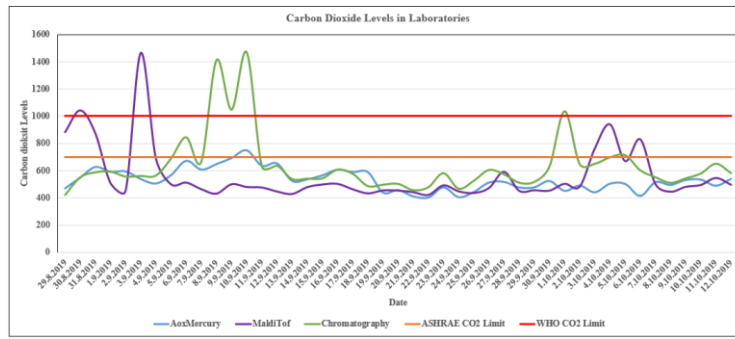


Fig. 6. The daily average CO₂ value between 29.08.2019 and 12.10.2019 in all laboratories.

B. Pre-Processing and Data Manipulation

Determination of Classes

The accepted reference values of important parameters that determine indoor air quality such as CO₂, CO, TVOC, PM_{2.5}, PM₁₀ have been determined by the institutions that are accepted worldwide such as WHO, Environmental Protection Agency (EPA), American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHARE). In this study, these reference values are used while classifying and labelling the data. However, while determining the limit values of parameters such as temperature and humidity, the past experiences of researchers who made analyses in other research and laboratories have been used. Although the level of light, which is the last parameter measured, is effective in many laboratory processes such as bacterial growth, no data indicating its impact on indoor air quality has been recorded. Generated classes and their limit values are shown in Table 1.

TABLE I
CLASSES AND LIMIT VALUES OF ENVIRONMENTAL PARAMETERS

	Excellent (5)	Good (4)	Moderate (3)	Poor (2)	Terrible (1)
T (°C)	18-19	17-18	16-17	<16	
	19-21	21-22	22-23	23-24	>24
Humidity	30-40	20-30	10-20	<10	
	40-60	60-70	70-80	80-90	>90
CO₂	<700	700-900	900-1100	1100-1300	>1300
TVOC	<40	40-70	70-100	100-150	>150
PM_{2.5}	<10	10-20	20-30	30-40	>40
PM₁₀	<20	20-40	40-60	60-80	>80
CO	<25	25-50	50-75	75-100	>100
Light	Nan	Nan	Nan	Nan	Nan

In many respected studies, generally, one parameter and two different classes are used, such as “Good” and “Poor” [40]. Since the overall purpose of this study is to find a suitable prediction algorithm for ontological sensor data, the situation for the algorithms to be selected is shaped to present a more complicated state; 5 different classes are defined for 7 parameters and the limit values are determined. The class label of an instance is identified by the parameter with the worst value of class among the parameters that make up that row. Table 2 shows how the class value of the row is determined.

When the instances are classified according to the aforementioned rule, it has been seen that 65% of the total of 3168 rows of data are at the desired level for the laboratory interior environment. However, in the remaining 35%, timely preparation of necessary action plans is vital for laboratory analysis results, and employee health. The experiments reveal that laboratory air quality is monitored at ideal ranges when there is no biological analysis and no human activity in the environment. Fig. 7 shows the number of individuals in each class.

Certain pre-processes are required to make logical inferences and obtain good conclusions on the data collected. Pre-processes such as removing noisy data, conveniently filling missing data, shift all parameter values to the same range (normalization) are absolutely necessary for determining a better prediction model. Pre-process operations performed before making estimates on the data and how they are applied are explained in detail below.

Missing Value Imputation

On the specified dates, approximately 25,920 data would be expected to have been saved to the Apache Jena Fuseki RDF database, though only 23,252 data have been recorded due to the malfunction of the devices operating in the system or human error. This number corresponds to approximately 90% of the data that should be recorded. It is important to fill the missing values with a reasonable approach, especially if the algorithms in effect that are sensitive to missing values such as Decision Tree (DT) and Random Forest (RF) are to be studied. In this manner, the missing 10% has been filled with the well-known and accepted methods, and data continuity was ensured.

In data mining, it is possible to deal with the missing value issue with different approaches, such as deleting the missing values, accepting the average of that feature as the standard, or accepting them zero. Deleting or statistically filling missing values causes bias and negative effects on the results. Therefore, unlike these approaches, inputting data can significantly improve the quality of the data set [37]. Recently, many studies have shown that filling missing values with classification approaches has positive effects on the output [13, 41, 42]. In our work, missing values are filled by utilizing a hybrid approach of the K-nearest neighbor (K-NN) algorithm and Decision Tree, and the quality of the data set is increased.

TABLE II
DETERMINING THE CLASS VALUES OF PARAMETERS AND ROWS

Temperature	Humidity	CO ₂	TVOC	PM _{2.5}	PM ₁₀	CO	Light	Nominal
22.93	54.16	534.55	20.86	10.66	12.85	27	74.63	Moderate
23.01	53.78	541.1	21.68	10.09	11.83	27	67.1	Poor
21.03	42.12	422	2.48	0.88	1.12	21.6	26	Good
20.99	42.2	417.45	1.71	1.32	1.38	21	4	Excellent
20.27	50.94	879.46	71.31	5.08	5.78	32.59	78.07	Moderate
20.31	50.94	554.24	23.08	4.67	5.73	32.8	76.56	Good
20.25	52.34	1348.59	142.37	7.58	8.96	37.28	28	Terrible
20.31	52.3	1223.55	128.47	7.79	9.22	34.65	28	Poor
19.66	52.25	1306.33	138.5	6.53	7.71	255.35	79.43	Terrible
19.59	55.33	407.04	0.28	3.42	3.73	22.57	26	Excellent

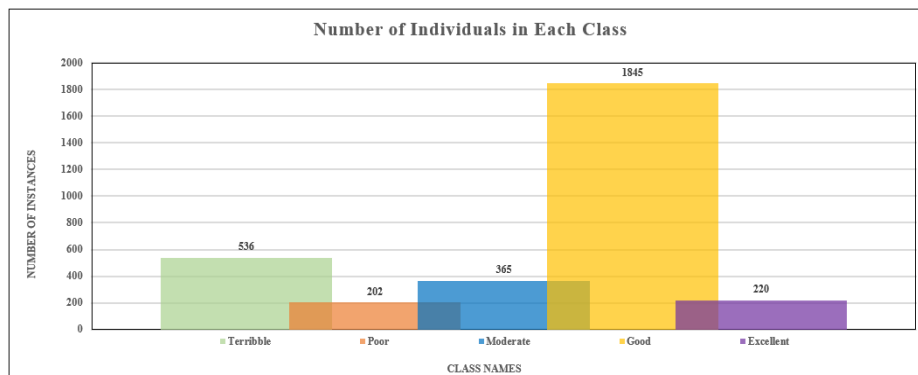


Fig. 7. Scatter graph of classes by row.

Outlier Detection

An Outlier can be defined as any observation different from other observations in the data set [43]. Outliers in the data collected by WSN can generally be caused by sensor measurement errors or some problems arising from data communication. Occasionally the outliers can be caused by human error. For example, if someone blows or touches the sensor in an environment where the temperature parameter is measured, this is a human error that causes the sensor value to deviate upwards. Both system-based and human-based errors cause the estimation to be biased and wrong. Therefore, analyzing the collected data and eliminating some inconsistent parts will increase the accuracy of the prediction.

There are some types of outlier detection approaches such as Probabilistic, Distance Based (cosine, Euclidean distance, etc.), algorithm-based (neighbor based, neural networks based, etc.). We evaluated outlier detection in two stages. First, the outlier data in each attribute has been found in itself and eliminated. During this process, the cosine distance approach, which is one of the distance-based outlier detections measures, is used, and a total of 10 observation data inconsistent with the other data have been deleted from each column. In the second step, after the class label of each instance (row) is assigned, outliers have been determined over this class label and eliminated. While determining outliers, the K-NN neighborhood approach has been used ($k=10$) and a total of 20 observation data eliminated.

Normalization

The measurement ranges, limit values, of each sensor used in this study are different. The measuring range is the total range that the instrument can measure under normal conditions. Table 3 shows the maximum and minimum values that can be measured by the sensors used in this study.

Absolute distance measuring methods such as Euclidean and Minkowski consider features in the same value ranges in the similarity calculation with equal importance. When using such distance measuring methods, calculating the similarity between instances without any pre-processing on the data set causes the feature with a large variance to have a high effect on the result [44]. In other words, the feature with large variance dominates the effect of other features on the result. It is called "feature domination". Moreover, the feature with high variance may not have a positive correlation with data in the same class, so it may not have the capability to parse data properly. In this case, the classification process might be misleading. To avoid feature domination; (i) all features are shifted to a certain interval. Normalization has significantly increased the performance of the classifiers used in this study. (ii) Cosine like similarity measures can be used that are not affected by the feature domination problem.

As demonstrated in Table 3, the values of some parameters can be between 0 and 100, while some parameter values may go up to 10,000. Therefore, it is certain that the prediction algorithms will decide according to the parameter with high values. In order to prevent this situation and to ensure that the parameters affect the estimation algorithm equally, all parameters have been shifted to the range of [0-1].

TABLE III
VALUE RANGES OF MEASURED PARAMETERS

Sensor	Parameter	Unit	Measurement Range
DHT22	Temperature	°C	-40 °C-125 °C (± 0.5)
DHT22	Humidity	% rh	0%-100% $\{\pm 2.5-5\}$
CCS-811	Carbon Dioxide	ppm	400-29,206 ppm
CCS-811	TVOC	ppb	0-32,768 ppb
Nova PM	Particular Matter 2.5	ppm	0.0-999.9 ppm
Nova PM	Particular Matter 10	ppm	0.0-999.9 ppm
MQ-7	Carbon Monoxide	ppm	10-10,000 ppm
LDR	Light Level	%	0%-100%

V. EXPERIMENTAL RESULTS

The results of classification algorithms on the aforementioned data set are presented in this section with different aspects. The results indicated in the figures and tables are the outcomes obtained for the test dataset. In order to reveal the achievements of algorithms, they have been run on the collected data set and it has been evaluated that the testbed established as a real-life case is sufficient for a fair evaluation of the classifiers. The algorithm performance tests have been performed on a computer with windows 10 operating system and equipped with Intel I7 7700HQ 2.8 GHz processor, 16 Gb DDR4 Memory, Nvidia Geforce Gtx 1050 video card.

When algorithm performance tests have been enforced on ontological sensor data, 70% percent of the data has been divided into the training set, and 30% percent test set. Indoor environmental parameters do not change rapidly at a dramatic pace. The deployed sensors tend to measure similar values in recent times and locations, meaning they are generally records of the same class. Therefore, it is crucial to be extremely careful when partitioning the dataset to minimize potential biases resulting from dataset partitioning. The dataset used is imbalanced. Working with k-fold on imbalanced datasets can lead to some challenges. If the records that make up the dataset are not randomly shuffled before applying k-fold, the test data may consist solely of records from a single class. The dataset includes temporal records, and if the dataset is randomly shuffled before applying k-fold, the temporality of the dataset may be affected, leading to an unfair evaluation. Due to the mentioned reasons, when creating the test data, records are selected randomly from different time periods, different classes, and different locations. The aim is to achieve a homogeneous distribution. 6 out of 9 machine learning algorithms evaluated are used with default parameter values. However, depths and the maximum number of tree parameters of RF, GBT, and DT algorithms negatively affect the time performances at their default values. Therefore, these parameters have been optimized for these algorithms, without much compromise on accuracy. The Maximal Depth and Number of Trees parameters are set to 10 in order to compete with other algorithms in terms of time.

All of the algorithms obtained acceptable accuracy values except Naive Bayes (NB) and Logistic Regression (LR). But the most successful algorithms in terms of accuracy among them are RF, Deep Learning (DL), and DT with the value 90%, 89%, 88% respectively. Therefore, it has been observed that these three algorithms are equally suited for this case. Generally, complexity and accuracy performance specify a trade-off in many cases, for this scenario the performance/complexity ratio of DT is better than others. The comparison of the accuracy percentages of the algorithms used in the case study is shown in Fig. 8.

The results obtained in terms of time comparison of the algorithms can be seen in Fig. 9. According to the results, we see that the most effective algorithms at the total time aspect are DT and Generalized Linear Model (GLM) methods, respectively. The biggest reason underlying the high speed of DT is the fast decision-making mechanism thanks to its tree structure. Also, DT doesn't need a large training set to get good results. GLM is a regression-based method and it is obvious and known that regression-based methods are effective especially in terms of running time. So it is not surprising that DT and GLM achieve the fastest scores. However, DT, NB, and RF algorithms have shown a tendency to learn faster. For this reason, the training time of these algorithms is the lowest. In addition, the duration of time spent in a training set with 1000 records are observed in the time graph in Fig. 9. According to this statistic, DT again gets the lowest score while DL gets the second place. This graphic demonstrates that the DL method has good scalability.

Fig. 10 shows the average correlations calculated by all models between labels and attributes. According to these correlations, the most important parameters affecting the result is PM₁₀, PM_{2.5}, and Temperature. While it is predictable that PM₁₀ and PM_{2.5} are active attributes, it is a surprise that the temperature is effective. However, the lectures in labs have increased the human presence and activity and the linear relationship of the temperature attribute with CO₂ has been caused by this situation.

The results in Fig. 10 revealed that the parameter of light does not have much effect on the results obtained however, it is an expected result. While setting the label value of each row in advance, it has been thought that the parameter of light would not affect the result and it is stated that it is not used in defining the line label.

In addition to the run time and accuracy comparisons of the selected algorithms, the amount of gain and loss is also an important parameter in the selection of the algorithm, especially in multi-class labelling. In a multi-class dataset, more acceptable it is for a predicted value to be in a class close to the real value than if in a class far from the true value is. The benefits and costs of the wrong and correct estimates are given in Table 4. Losses are represented as negative numbers while benefits or gains are represented as positive numbers.

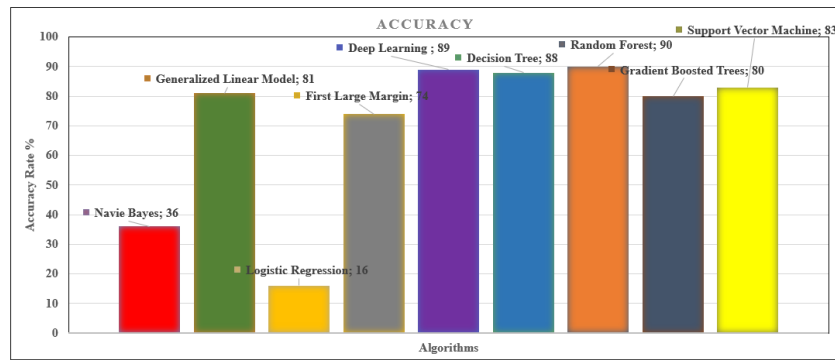


Fig. 8. Comparison of accuracy percentages of algorithms used in the case study.

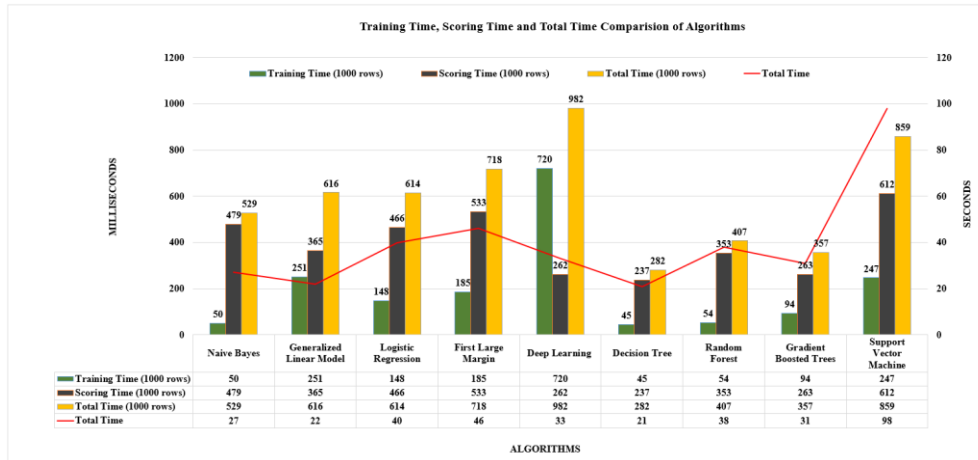


Fig. 9. Comparison of the training time, the scoring time, and the total time of algorithms used in the case study.

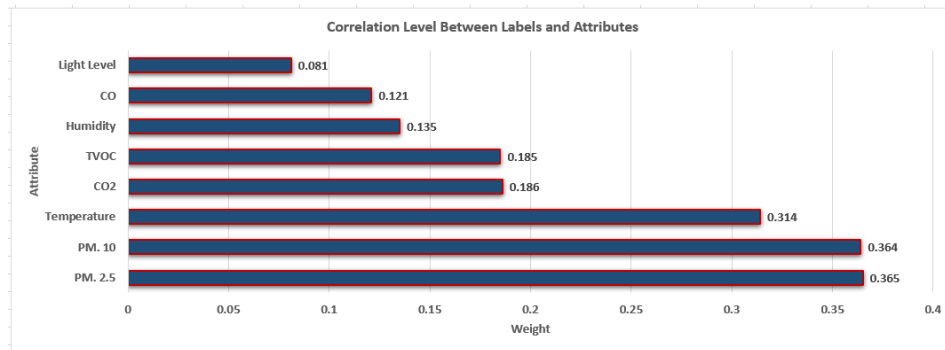


Fig. 10. The average correlations calculated by all models used between labels and attributes are seen.

For example, if the label value of an instance with an actual label value of Excellent is estimated as Excellent with any classifier, the prediction is correct and takes 1 as the gain point. On the other hand, if the classifier labelled the same Excellent instance as a Good, Moderate, Poor, or Terrible the classifier takes -1, -2, -3, -4 loss point respectively and this prediction becomes wrong. These loss points give the value of the wrong prediction. In some cases, it may be more beneficial to choose the best performing algorithm by looking at gain

rather than accuracy.

When the performances of the algorithms are compared in terms of gain, it is seen that the sum of the costs of NB and LR algorithms is negative, while the remaining algorithms are positive. When the performances of algorithms are assessed via gain metric, it is seen that the algorithms that give the best results in parallel with their accuracy rates are RF, DL, and DT. A comparison of algorithm performances in terms of gain is given in Fig. 11.

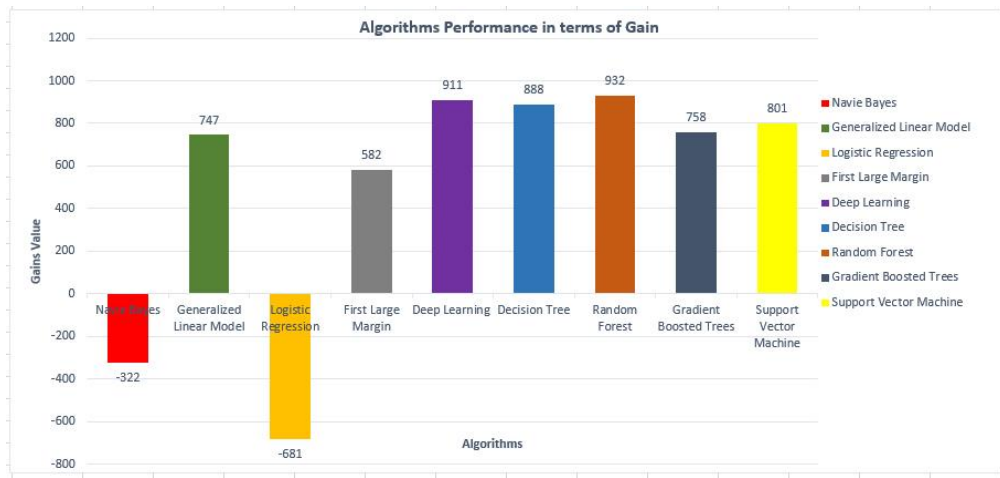


Fig. 11. Comparison of algorithms used in the proposed system from the perspective of gain.

TABLE IV
COST MATRIX REFERENCED WHEN COMPAIRING THE GAIN PERFORMANCE OF ALGORITHM USED

Cost Matrix	True Terrible	True Poor	True Moderate	True Good	True Excellent
Predicted Terrible	1	-1	-2	-3	-4
Predicted Poor	-1	1	-1	-2	-3
Predicted Moderate	-2	-1	1	-1	-2
Predicted Good	-3	-2	-1	1	-1
Predicted Excellent	-4	-3	-2	-1	1

VI. CONCLUSION AND FUTURE WORK

In recent years, sensor-based systems have rapidly spread to all areas of daily life as a result of the physical minimization of sensors in size, enabling the use in every field, the developments in the academic community, and the decrease in their prices. The intensive use of sensor and sensor based systems in every field has caused an exponential increase in sensor data in the internet environment. However, the heterogeneous nature of the sensor data makes it difficult to manage them under a single infrastructure. In addition, the absence of a common framework for the representation of sensor data makes it difficult for the machines to be understood and interpreted. Although a syntactic relationship has been established between sensor data in studies conducted so far, this is insufficient to make meaningful inferences from sensor data.

Semantic Sensor Web technology has been suggested and used by many researchers to address all these problems. Creating semantic relationships instead of establishing syntactic relationships between sensor data will provide more meaningful inferences. In addition, sensor data must be encoded in languages that machines can understand and interpret, such as RDF and OWL. Each sensor data should be represented by URIs and it should be easier for data consumers to reach it. In the first step of this study, a different model has been created by using the SSN framework to manage the data collected from different platforms, different environments, and different sensors under the same infrastructure. In the second step of the proposed study, in

order to establish a proactive system design, some traditional and state-of-art prediction algorithms on ontological sensor data are tested and compared by using data from this model. When the values obtained by running the algorithms on the collected sensor data are compared, it is seen that the most effective algorithms are RF and DT in terms of run time, accuracy and gain.

The proposed model can be combined with different domains, different platforms, and different systems to expand its scope in future studies. With this extended model, sensor data can be used to make a common inference. Although the proposed sensor ontology associates the data semantically, the complexity of the semantic techniques often causes an increase in processing times. A new model that includes minimum concepts to ensure that the proposed semantic systems respond in a reasonable acceptable time to data consumers can be created. Object properties and data properties can be used within the scope of the minimum concept. Thus, the triple number in the RDF database is reduced and the system can be more efficient.

The ontological sensor data framework developed within the scope of the study, while providing a range of advantages such as semantic enrichment of data and reusability, is also considered to have some weaknesses. Undoubtedly, the major disadvantage of ontological datasets is low coverage and high complexity. Creating ontological datasets can be a complex process, and developing a comprehensive ontology for a subject may take time. The ontological dataset created for any subject for the first time must adapt to the changes occurring in that field. This situation necessitates making additional updates to the ontological dataset over time. In addition to all these problems, excessively enriching the dataset semantically will lead to unnecessary overloading of the dataset. Therefore, the decision on the extent to which raw sensor data should be enriched needs to be made considering the cost-benefit ratio.

Acknowledgments

The authors would like to thank the Scientific, Industrial and Technological Application and Research Center of Bolu Abant İzzet Baysal University for utilization of MaldiToF laboratory, AoxMercury laboratory, and Chromatography laboratory, as real-world use-case in proposed sensor ontology.

Data availability

All data of 8 measurements collected over 45 days using 5 different sensors from 3 different laboratories are in the links:

link1: <https://doi.org/10.4121/14805960.v1>

link2:

https://figshare.com/articles/dataset/Labs_Sensor_Data/14742858

REFERENCES

- [1] L. Bermudez, E. Delory, T. O'Reilly and J. Del Rio Fernandez, "Ocean observing systems demystified", *MTS/IEEE Biloxi - Mar. Technol. Our Futur. Glob. Local Challenges, Ocean*, 2009, pp. 1–7.
- [2] S. Abd Hakim, K. Tarigan, M. Situmorang, and T. Sembiring, "Synthesis of Urea Sensors using Potentiometric Methods with Modification of Electrode Membranes Indicators of ISE from PVA-Enzymes Coating PVC-KT p CIPB", *J. Phys. Conf. Ser.*, vol. 1120, no. 1, 2018.
- [3] A. Sheth, "Interoperating Geographic Information Systems", *Interoperating Geogr. Inf. Syst.*, pp. 5–30, 1999.
- [4] F. Wang, L. Hu, J. Zhou, J. Hu and K. Zhao, "A semantics-based approach to multi-source heterogeneous information fusion in the internet of things", *Soft Comput.*, vol. 21, no. 8, pp. 2005–2013, 2017.
- [5] M. Arooj, M. Asif and S. Zeeshan, "Modeling Smart Agriculture using SensorML", *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 5, pp. 0–6, 2017.
- [6] A. Haller *et al.*, "The modular SSN ontology: A joint W3C and OGC standard specifying the semantics of sensors, observations, sampling, and actuation", *Semant. Web*, vol. 10, no. 1, pp. 9–32, 2018.
- [7] J. Liu, Y. Li, X. Tian, A. K. Sangaiah and J. Wang, "Towards semantic sensor data: An ontology approach", *Sensors (Switzerland)*, vol. 19, no. 5, 2019, pp. 1–21.
- [8] H. K. Patni and C. A. Henson, "Linked Sensor Data", 2010, pp. 362–370.
- [9] A. N. U. Armin Haller, S. B. Krzysztof Janowicz, University of California, C. Simon Cox, T. U. of B. Danh Le Phuoc, A. N. U. Kerry Taylor, and É. N. S. des M. de S.-É. Maxime Lefrançois, "Semantic Sensor Network Ontology—W3C," 2011. [Online]. Available: <https://www.w3.org/TR/2017/REC-vocab-ssn-20171019/>. [Accessed: 20-May-2021].
- [10] P. Barnaghi *et al.*, "Semantic Sensor Network XG Final Report", 2017.
- [11] J. P. Calbimonte, H. Jeung, O. Corcho and K. Aberer, "Enabling query technologies for the semantic sensor web", *Int. J. Semant. Web Inf. Syst.*, vol. 8, no. 1, 2012, pp. 43–63.
- [12] S. Avancha, C. Patel and A. Joshi, "Ontology-driven adaptive sensor networks", *Proc. MOBIQUITOUS 2004 - 1st Annu. Int. Conf. Mob. Ubiquitous Syst. Netw. Serv.*, 2004, pp. 194–202.
- [13] M. Chen, J. Zhou, G. Tao, J. Yang and L. Hu, "Wearable affective robot", *IEEE Access*, vol. 6, 2018, pp. 64766–64776.
- [14] L. Hu, J. Yang, M. Chen, Y. Qian and J. J. P. C. Rodrigues, "SCAI-SVSC: Smart clothing for effective interaction with a sustainable vital sign collection", *Futur. Gener. Comput. Syst.*, vol. 86, 2018, pp. 329–338.
- [15] H. Rathore, A. Al-Ali, A. Mohamed, X. Du and M. Guizani, "DLRT: Deep learning approach for reliable diabetic treatment", *IEEE Glob. Commun. Conf. GLOBECOM 2017 - Proc.*, vol. 2018, 2017, pp. 1–6.
- [16] A. A. Sarangdhar, P. V. R. Pawar and A. B. Blight, "Machine Learning Regression Technique for using IoT", *Int. Conf. Electron. Commun. Aeronaut. Technol. ICECA 2017*, pp. 449–454.
- [17] S. S. Patil and S. A. Thorat, "Early detection of grapes diseases using machine learning and IoT", *Proc. - 2016 2nd Int. Conf. Cogn. Comput. Inf. Process. CCIP 2016*, pp. 7–11.
- [18] I. U. Din, M. Guizani, J. J. P. C. Rodrigues, S. Hassan and V. V. Korotaev, "Machine learning in the Internet of Things: Designed techniques for smart cities", *Futur. Gener. Comput. Syst.*, vol. 100, 2019, pp. 826–843.
- [19] N. J. Patel and R. H. Jhaveri, "Detecting Packet Dropping Misbehaving Nodes using Support Vector Machine (SVM) in MANET", *Int. J. Comput. Appl.*, vol. 122, no. 4, 2015, pp. 26–32.
- [20] J. Canedo and A. Skjellum, "Using machine learning to secure IoT systems", *2016 14th Annu. Conf. Privacy, Secur. Trust. PST 2016*, pp. 219–222.
- [21] I. Kotenko, I. Saenko, F. Skorik and S. Bushuev, "Neural network approach to forecast the state of the Internet of Things elements", *Proc. Int. Conf. Soft Comput. Meas. SCM 2015*, pp. 133–135.
- [22] M. Bermudez-Edo, T. Elsaleh, P. Barnaghi and K. Taylor, "IoT-Lite: a lightweight semantic model for the internet of things and its use with dynamic semantics", *Pers. Ubiquitous Comput.*, vol. 21, no. 3, 2017, pp. 475–487.
- [23] I. Yang, "Design and Implementation of e-Health System Based on Semantic Sensor Network Using", 2018.
- [24] C. Kuster, J. L. Hippolyte and Y. Rezgui, "The UDSA ontology: An ontology to support real time urban sustainability assessment", *Adv. Eng. Softw.*, vol. 140, 2020, pp. 102731.
- [25] C. Wang, Z. Chen, N. Chen and W. Wang, "A hydrological sensor web ontology based on the SSN ontology: A case study for a flood", *ISPRS Int. J. Geo-Information*, vol. 7, no. 1, 2018.
- [26] S. Ali, S. Khusro, I. Ullah, A. Khan and I. Khan, "SmartOntoSensor: Ontology for Semantic Interpretation of Smartphone Sensors Data for Context-Aware Applications", vol. 2017, 2017.
- [27] J. Adeleke, D. Moodley, G. Rens and A. Adewumi, "Integrating Statistical Machine Learning in a Semantic Sensor Web for Proactive Monitoring and Control", *Sensors*, vol. 17, no. 4, 2017, pp. 807.
- [28] A. C. Onal, O. Berat Sezer, M. Ozbayoglu and E. Dogdu, "Weather data analysis and sensor fault detection using an extended IoT framework with semantics, big data, and machine learning", *Proc. - 2017 IEEE Int. Conf. Big Data*, 2017 pp. 2037–2046.
- [29] I. A. Al-Baltah, A. A. A. Ghani, G. M. Al-Gomaei, F. A. Abdulrazzak and A. A. A. Kharusi, "A scalable semantic data fusion framework for heterogeneous sensors data", *Journal of Ambient Intelligence and Humanized Computing*, 2020, pp. 1–20.
- [30] C. Kuster, J. L. Hippolyte and Y. Rezgui, "The UDSA ontology: An ontology to support real time urban sustainability assessment", *Advances in Engineering Software*, vol. 140, 2020, 102731.
- [31] A. A. Sarangdhar and V. R. Pawar, "Machine learning regression technique for cotton leaf disease detection and controlling using IoT", In *2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)* vol. 2, 2017, pp. 449–454.
- [32] World Health Organization, "Global status report on noncommunicable diseases 2014", (No. WHO/NMH/NVI/15.1), World Health Organization, 2014.
- [33] A. Donkers, D. Yang, B. de Vries and N. Baken, "Semantic web technologies for indoor environmental quality: A review and ontology design", *Buildings*, vol. 12, no. 10, 2022.
- [34] F. Desimoni, S. Ilarri, L. Po, F. Rollo and R. Trillo-Lado, "Semantic traffic sensor data: The TRAFair experience", *Applied Sciences*, vol. 10, no. 17, 2020.
- [35] S. Domínguez-Amarillo, J. Fernández-Agüera, S. Cesteros-García and R. A. González-Lezcano, "Bad air can also kill: residential indoor air quality and pollutant exposure risk during the COVID-19 crisis", *International Journal of Environmental Research and Public Health*, vol. 17, no. 19, 2020.
- [36] R. Mumtaz *et al.*, "Internet of things (IoT) based indoor air quality sensing and predictive analytic—A COVID-19 perspective", *Electronics*, vol. 10, no. 2, 2021.
- [37] Ö. Aktaş, M. Milli, S. Lakestani and M. Milli, "Modelling sensor ontology with the SOSA/SSN frameworks: a case study for laboratory parameters", *Turkish Journal Of Electrical Engineering And Computer Sciences*, vol. 28, no. 5, 2020, pp. 2566–2585.
- [38] M. A. Musen, "The protégé project: a look back and a look forward", *AI matters*, vol. 1, no. 4, 2015, pp. 4–12.
- [39] Apache Software Foundation, "'Apache Jena' A free and open source Java framework for building Semantic Web and Linked Data applications", [Online]. Accessed on September 11, 2021. <https://jena.apache.org/documentation/fuseki2/index.html>.
- [40] J. A. Adeleke, D. Moodley, G. Rens and A. O. Adewumi, "Integrating statistical machine learning in a semantic sensor web for proactive monitoring and control", *Sensors*, vol. 17, no. 4, 2017, 807.
- [41] N. Z. Abidin, A. R. Ismail and N. A. Emran, "Performance analysis of machine learning algorithms for missing value imputation", *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 6, 2018.
- [42] N. D. Darryl and M. M. Rahman, "Missing value imputation using stratified supervised learning for cardiovascular data", *J Inform Data Min*, vol. 1, no. 13, 2016.

- [43] V. Barnett and T. Lewis, "Outliers in statistical data", John Wiley & Sons, Chichester, 1994.
- [44] A. K. Jain, M. N. Murty and P. J. Flynn, "Data clustering: a review", *ACM computing surveys (CSUR)*, vol. 31, no. 3, 1999, pp. 264-323.

BIOGRAPHIES



MEHMET MİLLİ Karşıyaka, İZMİR in 1984. He graduated from the Computer Engineering Department of Mersin University. He received M.Sc. and Ph.D. degrees in Computer Engineering from the University of Dokuz Eylül in 2016 and 2021 respectively. His research interests include embedded systems, artificial intelligence, data mining and sensor networks.



ÖZLEM VARLIKLAR Adana, in 1981. She received the B.Sc., M.Sc., and Ph.D. degrees in computer engineering from the University of Dokuz Eylül, Izmir, Turkey, in 2003, 2005, and 2010, respectively. Since 2004, she has been with the Department of Computer Engineering, University of Dokuz Eylül, where she is currently assistant professor. Her research interests include artificial intelligence, computer software, decision support systems and natural language processing.



MUSA MİLLİ Karşıyaka, İZMİR in 1984. He graduated from the Computer Engineering Department of Sakarya University. He received M.Sc. and Ph.D. degrees in Computer Engineering from the University of Ege, İzmir, Turkey in 2013 and 2019 respectively. He has been with the Department of Computer Engineering of Turkish Naval Academy, National Defense University, Tuzla, İSTANBUL. His research interests include artificial intelligence, data mining, machine learning, information retrieval, big data and network security.



SANAZ LAKESTANI Abadan, IRAN in 1975. She graduated from Sehit Behesti University in 1997. She received the M.Sc. from the University of Islami Azad in 2001 and Ph.D in Environmental Engineering Department of Hacettepe University in 2015. Her research interests include air pollution and control, environmental impact assessment and risk management.