International Journal of Computing and Digital Systems
2025, VOL. 17, NO.1, 1-12

http://dx.doi.org/10.12785/ijcds/1571028689

# An Ontology Alignment based on Machine learning for Integration of Patient Health Data

**Nidhi Gupta[1], Pawan Kumar Verma[1,*], Sundeep Raj[2], Anushree[3], Nitin Rakesh[4] and Monali Gulhane[4]**

[1]*School of Engineering and Technology, Sharda University, Greater Noida, Uttar Pradesh, India;*
[2]*KIET Group of Institutions, Ghaziabad, Uttar Pradesh, India;*
[3]*GLA University, Mathura, Uttar Pradesh, India;*
[4]*Symbiosis Institute of Technology, Nagpur Campus, Symbiosis International (Deemed University), Pune, Maharashtra, India;*

*Received 8 May 2024, Revised 29 January 2025, Accepted 29 January 2025*

**Abstract:** The integration of patient data is crucial in healthcare informatics. It involves organizing and integrating heterogeneous health data from various Electronic Health Records (EHRs). Attribute alignment is a fundamental step in data integration. It involves mapping data attributes across different datasets. Most of the data maintained in EHRs does not follow standard terminologies in healthcare. Therefore, it becomes difficult to integrate patient health data from diverse data sources for generating historic medical records. The research work carried out overcomes this problem by developing a vital sign ontology using OpenEHR health standards. It helps to map the vital signs observations of the patients from its proprietary sources uniformly. The work also leverages the power of supervised learning algorithms to automate the mapping of different health datasets to the proposed ontology. The approach is evaluated on patient health datasets, considering both standard and non-standard datasets. The research work employs different machine learning algorithms, such as Support Vector Machine (SVM), Naive Bayes, Logistic Regression, k-nearest neighbor (KNN), AdaBoost, and Neural network, in order to evaluate the best algorithm for the proposed approach. The evaluation results conclude that Naive Bayes exhibits the highest accuracy, with minimum misclassification rate, in both the training and validation phases for automatically mapping the health datasets with the proposed ontology.

**Keywords:** Electronic Health Record, machine learning, OpenEHR, Ontology alignment, interoperability, schema mapping

## 1. Introduction

The integration of large volumes and a variety of patient data is a major concern in the healthcare domain. Data integration requires data interoperability. Healthcare data comes from diverse sources, including hospitals, clinics, laboratories, wearable devices, and patient-generated data. It has resulted in interoperability issues, thus making data integration a challenging task. Semantic data integration aims to preserve the meaning of the data between various data sources. It provides meaningful data integration. Ontology plays an integral role in performing semantic data integration [1]. It aids in the improvement of semantic interoperability [2]. Ontology-based data integration aims to bridge these differences by mapping data elements to ontology concepts. Ontologies provide a structured and standardized way to represent and describe medical concepts. It aids in harmonizing and making sense of diverse health data for a range of uses, such as clinical decision support, research endeavors, and population health management. Common healthcare ontologies include SNOMED-CT, LOINC, and UMLS (Unified Medical Language System) [3]. These ontologies provide a common vocabulary and semantic framework for data integration.

The Ontology incorporates domain information via classes, objects, and data properties. The classes represent a concept of a domain. It is defined in the form of a hierarchy. A class may have many sub-classes. The relationships provide the association between the concepts. The attributes describe various features of the concept, and the elements in the ontology are represented by its instances. Integrating health data from various sources provides a complete view of a patient's medical history, leading to more accurate diagnoses and personalized treatment plans. The healthcare providers can identify trends, predict outcomes, and take preventive measures, ultimately improving patient outcomes. It also facilitates in medical research. However, integration of patient data has significant challenges such as data silos, interoperability, data privacy and legal issues. Integrating machine learning with ontology alignment creates a strong framework for clinical decision support by integrating standardized data with advanced analytics and improved accuracy. The integration of different data sources needs to align their schemas to the Ontologies. In healthcare sector also, the health records of the patients

*E-mail address: nidhi0208@gmail.com, abes.pawan@gmail.com (*Corresponding Author), sundeepraj1@gmail.com, anushree.gla@gla.ac.in, nitin.rakesh@gmail.com, monali.gulhane4@gmail.com*

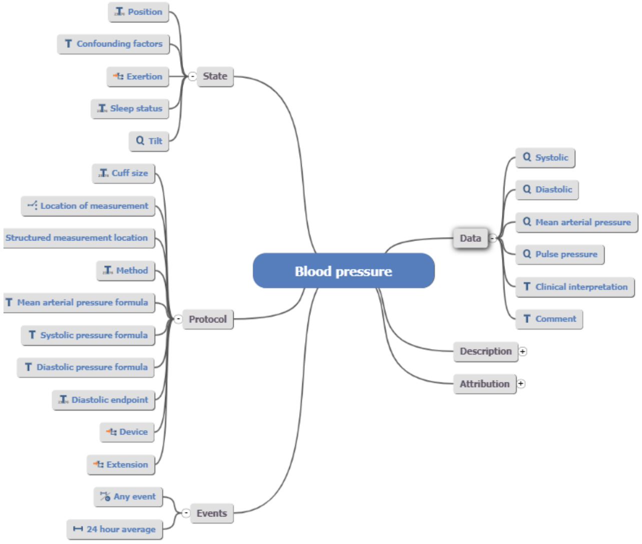

Figure 1. Mind map diagram of blood pressure concept by OpenEHR [4]

are maintained by various health providers. The patient EHR employs diverse vocabularies for the representation of a concept. Thus, there is a need to develop an Ontology that represents the global schema for data integration, and map the data sources to the global schema. The research work carried out proposed a Vital Sign Ontology (VSO) to provide a unified view of patient vital signs data sources.

The work utilizes the openEHR [4] health standard for development of ontology. It is an open standard that outlines the management, store, retrieval, and exchange of health data in electronic health records. It aims to standardization of record to achieve universal interoperability among all electronic health data.

OpenEHR provides an archetype model of a clinical domain for describing clinical knowledge. An archetype is a specification of different clinical concepts such as ECG result, Blood Pressure, etc. The figure 1 shows the mind map of OpenEHR archetype. The objective of this research work is to design a Vital Signs Ontology (VSO) based on openEHR health standard for schema alignment. A schema mapping approach is applied to find the best machine learning algorithm to map different data sources with the proposed VSO ontology. Schema mapping is a vital technique in health informatics that affects data integration, interoperability, and overall healthcare quality. It plays an important role in transforming and normalizing data to ensure consistency across different sources. This research addresses the problem of schema mapping and alignment of health data from different sources while ensuring the data privacy.

## 2. Background

In the healthcare sector, data integration involves consolidating a large variety and volume of data from diverse sources such as hospitals, clinics, and wearable devices. The integration of these data sources is crucial for enhancing clinical decision-making and patient care.

### A. *Challenges of data interoperability*

Data interoperability within healthcare presents significant challenges, primarily due to the disparate nature of the systems and the lack of uniform standards. These challenges hinder the effective exchange and utilization of critical health information across different platforms.

### B. *The role of semantic data integration*

Semantic data integration is a process in which meaning is preserved across heterogeneous systems, allowing for the uniform understanding and use of information. It is the key mechanism to address the issues of interoperability through the delivery of a context-rich integrated data environment.

### C. *Ontologies in healthcare*

Ontologies are a key component of semantic data integration since they provide a structured and standardized way to represent medical concepts, hence supporting harmonization of very diverse health data. Other common healthcare ontologies will be SNOMED-CT, LOINC, and UMLS for a common understanding among different systems.

### D. *Development of a Vital Sign Ontology (VSO)*

The research focuses on developing a Vital Sign Ontology based on the openEHR standard. The VSO intends to standardize the representation of vital signs data, align various sources of data to a global schema, and facilitate the effective integration of healthcare information.

### E. *Research objectives*

The research is looking for better schema alignment and the search for viable machine learning algorithms that will perform the mapping of a set of data sources onto VSO. This approach leverages the openEHR archetype model, characterized by rich descriptions for different clinical concepts. Specifically, two important points of research are as follows:

1) *Schema alignment:* Align the data structure among the various sources of health information into one and a common unified schema, in this context being the VSO, which is aligned using the standard definitions from the openEHR archetype model.
2) *Machine learning mapping:* The identification of the most suitable machine learning techniques to be used for mapping various data sources to the unified VSO schema with accuracy.

## 3. Mathematical Representation

### A. *Symbols definition*

Let us define the symbols used in the formulation:

- $S$ - the collection of source schemas from multiple data sources.

- $D$ - the data set collected from these sources.
- $V$ - the schema of the VSO.
- $A$ - a set of openEHR archetypes, which define the structure and semantics that $V$ is supposed to meet.
- $f$ - the function of the map done by the machine learning algorithm, mapping data $D$ from schema $S$ to $V$.
- $\theta$ - the parameters of the machine learning model that is utilized in $f$.

### B. *Formula for mapping function*

The mapping function $f$ is defined by:

$$f(D;\theta) = \hat{V} \tag{1}$$

where $\hat{V}$ is the predicted alignment of data $D$ according to the VSO schema $V$, based on the parameters $\theta$.

### C. *Objective function to optimize*

The objective of the research is to find the optimal parameters $\theta$ that minimize the loss function $L$, which measures the discrepancy between the aligned data $\hat{V}$ and the VSO schema $V$:

$$\theta^* = \arg\min_{\theta} L(V, \hat{V}) \tag{2}$$

where $L$ could be a function that measures how well $\hat{V}$ aligns with $V$, incorporating various factors like the fidelity of mapping, the semantic accuracy based on $A$, and possibly other domain-specific criteria.

## 4. Related Work

Several ontology-based solutions for data exchange and integration among various clinical data sources has been proposed in the literature. Kock-Schoppenhauer et al. [] performed integration of clinical data at heterogenous isolated system using ontology and semantic web technologies. Peng et al. [5] performed an ontology-based solution for integration of heterogeneous health services and data captured in home environment. Many studies involve transformation of healthcare data to standard terminologies to enable its exchange. Kiourtis et al. [6] performs ontology alignment to map the health care data in RDF format to HL7 FHIR. Peng et al. [5] proposed an Ontology using HL7 FHIR interoperability standard to integrate health data sources with various other web-based health services. Cimmino et al. [7] provides semantic interoperability of various IoT devices through ontology mappings. It represents the specifications of these devices in RDF data format and provides SPARQL query interface to discover and access IoT devices. Gupta et al. [8] performed the machine learning approach to map the schema of different data sources. Frid et al. [9] uses the ontologies to consolidate the clinical data for clinical research. The traditional methods of schema mappings are rule based, lexical based etc. These techniques are simple and easy to implement. They are highly accurate for specific domains where rules are well-defined. However, these techniques are inflexible and difficult to maintain as ontologies evolve. They are prone to false positives due to homonyms and synonyms. Machine learning offers a transformative approach to addressing the limitations of traditional methods in both EHR data integration and ontology alignment. Machine learning algorithms are capable of efficiently processing large amounts of data, allowing for scalable and automated integration and alignment operations. They can continuously learn and adapt to new data, resulting in increased accuracy over time. Ontology-based solutions for integration of health data are widely studied in the past. The table I represents the related studies on ontology alignment and data integration. However, the work performed in the past does not define an ontology to record the basic vital signs of the patients such as Blood Pressure, temperature, pulse pressure, etc. These vital signs observations are the first step of any clinical evaluation and are the prime indicator of patient health readings. The research work carried out proposed a vital sign ontology according to the OPENEHR health standard for integrating patient data.

## 5. Methodology

The work carried out presents the methodology of VSO ontology and thereafter it maps the different sources to the proposed ontology using machine learning algorithm.

### A. *Vital sign ontology development*

The development of an ontology involves identification of concepts, classes and data properties. The work carried out developed an ontology for the concept of vital signs of a human body. The data properties of the ontology are designed in accordance with the data properties of the OpenEHR health standard archetype for different concepts of vital signs. It provides the global view of underlying data sources for data integration.

Vital signs indicate the status of the essential functions of the body. The vital signs are regularly monitored by health professionals for assessing the health of a person. The OpenEHR standard provides a template that contains the various data elements of a vital sign's concepts. There are four prime vital signs of the human body: Blood Pressure, pulse pressure, body temperature and respiratory rate.

The Blood pressure [19] is the measurement of arterial blood pressure of an individual. The concept of blood pressure is described in four different attributes accordance to OpenEHR health standard. These are Systolic (sys), Diastolic (dys), Mean Arterial Pressure (MAP) and Pulse Pressure (PP). Systolic is a measure the maximum arterial blood pressure. It is the contraction phase of the cardiac cycle. Diastolic is measures the minimum arterial blood pressure. It is the rest phase of the cardiac cycle. MAP is the average arterial pressure that occurs over the entire course of the cardiac cycle of the heart, and PP is the difference between Systolic and Diastolic observation. The other prime vital signs considered in this research are Body temperature

TABLE I. Related studies on ontology alignment and data integration

| Reference | Year | Methodology | Key Findings |
|---|---|---|---|
| Mitra et al. [10] | 1999 | Rule based approach written in first order logic | Support name matching and structural matches on is-a hierarchy |
| Wen et al. [11] | 2000 | Uses schema design information as meta-data for training neural network | Uses attribute cluster to produces similarity score to match it attribute in another schema |
| Madhavan et al. [12] | 2001 | Linguistic (attribute name) and structural (context) similarity | Performs Schema based mapping using name, data type and constraints of data sources. |
| Fagin et al. [13] | 2009 | It is a constraint-based approach. It uses the constraints to create a query on source schema and create a query on target schema. The schema is mapped using value correspondence | The approach is used for relational databases with constraints defined on schema value |
| Knoblock et al. [14] | 2012 | Uses Probabilistic and Rule based approach | Provides mapping between data sources and Ontology. It computes average probability of occurrence of an attribute to target class |
| Birgersson et al. [15] | 2016 | Approach used for mapping XML data formats using xpath. | Faced problems such as mapping of new unseen xpath, incorrect mapping with similar attribute name but different values. |
| Rouces et al. [16] | 2016 | It defines the heuristics and uses linguistic approach. | It finds the complex relationship between arbitrary linked datasets and mediated schema. |
| Rajkomar et al. [17] | 2018 | Scalable and accurate deep learning with electronic health records | Demonstrated the scalability and accuracy of deep learning models in predicting multiple medical events using EHR data in real-world clinical settings. |
| Xu et al. [18] | 2022 | Deep learning for EHR | Comprehensive coverage of use of deep learning for EHR |

(temp) [20], Pulse or Heart rate (rate) [21], and Respiration rate (resp) [22].

A variety of tools are available for Ontology development, such as Hozo [23], Swoop [24], etc. Protege is one of the open-source and widely used ontology development editors [25], [26]. It provides a graphical user interface for designing an ontology. It facilitates the building of an Ontology in various data formats such as OWL, XML, RDFS, etc. It provides a plug-and-play environment, which assists the fast development of an application. It is supported by a large community of users such as academic and business communities .

The proposed Vital Sign Ontology design consists of a class called Observation. It represents the measurements of various vital signs of a patient. The data properties of the observation class are the four main vital signs of the human body. These are Blood Pressure, Temperature, Heart rate, and Respiratory rate. The data properties of an ontology represent a relation that relates an entity to the datatype literals, such as date, number or a string. The data properties of the vital sign observation class are designed in accordance with the data provided by the OpenEHR standard. The graphical representation of VSO in Protege is shown in figure 2.

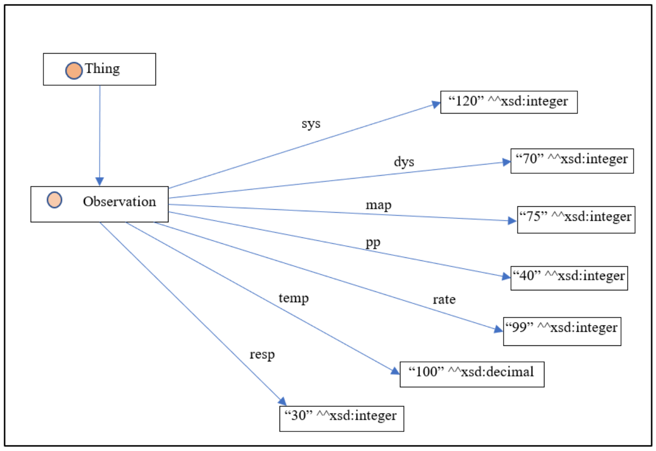

Figure 2. The graphical representation of proposed vital sign ontology showing classes and data properties

The VSO in Protege is implemented with two different classes: Thing and Observation.

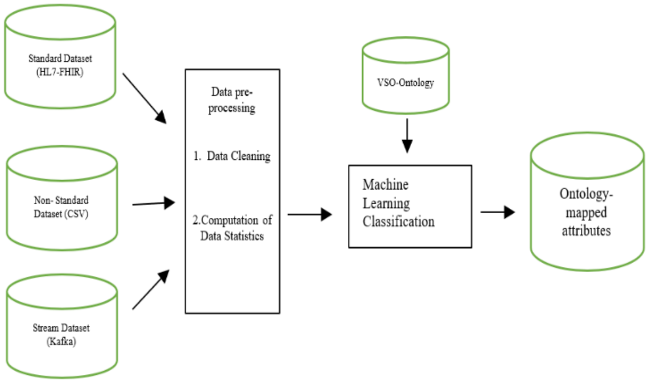
Standard Dataset (HL7-FHIR)
Non- Standard Dataset (CSV)
Stream Dataset (Kafka)
Data pre-processing
1. Data Cleaning
2.Computation of Data Statistics
VSO-Ontology
Machine Learning Classification
Ontology-mapped attributes

Figure 3. Architecture of SMOD approach

1) *Thing class:* The Thing class is a standard class in Protégé. All the other classes in the Protégé are defined under Thing class. It acts as a superclass.
2) *Observation class:* This class represents various vital sign observations of the patients. It has various data properties such as sys, dys, pp, map, rate, temp, and resp. The data properties relate the observation class with the individual data instances.

## B. Schema mapping

The schema mapping is performed by mapping the data source attributes to the corresponding ontology-mapped attributes. The proposed approach uses the SMOD [8] approach to map the proposed vital sign ontology. Figure 3 depicts the architecture of SMOD. The approach has three different phases.

### 1) Data extraction

This layer retrieves the data from different sources that encompass a wide range of stored data types. The data sources may be from three different category of data source, that are Standardized data, non-Standardized data and Data Streams. The Standardized data is the data that is stored and maintained in the pre-defined data format. The Fast Health Interoperability Resources (FHIR) is one of the the standard data format created by Health Level Seven(HL7) for easy exchange of EHR among the sources, whereas non-Standardized data is created by different organizations in their proprietary data format. This data may differ in their schema and data formats, and Data streams are flow of data generated by the data sources. It represents the data flow from producer to consumer in real time. The data stream can be divided into multiple windows. A record in a stream consists of the key, value and timestamp. The data is pre-processed and fetched using wrappers from their respective sources.

### 2) Pre-processing

The pre-processing is done to prepare the data for classification. It involves two key steps.

1) *Data cleaning, PCA and Feature Selection:* It involves identification and removal of anomalies of the datasets. It checks the data sources for missing values, noisy and incorrect data, along with data cleaning, the proposed model implements principal component analysis(PCA) and feature selection in the preprocessing.

#### Data Standardization

To ensure each feature has zero mean and unit variance, standardize the data using:

$$Z = \frac{X - \mu}{\sigma} \tag{3}$$

where $X$ is the original data, $\mu$ is the mean, and $\sigma$ is the standard deviation of each feature.

#### Covariance Matrix Computation

The covariance matrix of the standardized data is computed as follows:

$$\text{Cov(Z)} = \frac{1}{n-1} Z^T Z \tag{4}$$

where $Z^T$ is the transpose of $Z$, and $n$ is the number of data points.

#### Eigenvalue Decomposition

Compute the eigenvalues and eigenvectors of the covariance matrix. The eigenvectors represent the directions of maximum variance, and the eigenvalues represent the magnitude of the variance in those directions.

#### Select Principal Components

Sort the eigenvectors by decreasing eigenvalues and choose the top $k$ eigenvectors to form a projection matrix $P$.

#### Transform the Original Data

The transformed data $T$ in the new feature space defined by the principal components is given by:

$$T = ZP \tag{5}$$

#### L1 Regularization (Lasso)

The objective function in Lasso regularization is formulated as:

$$\min_{\beta} \left\{ \frac{1}{2n} ||Y - X\beta||^2 + \lambda ||\beta||_1 \right\} \tag{6}$$

where $Y$ is the output vector, $X$ is the feature matrix, $\beta$ are the coefficients, $\lambda$ is the regularization parameter, and $||\beta||_1$ is the L1 norm of the coefficients, encouraging sparsity.

#### *Adjusted Standardization Formula*

To enhance the robustness of the model against outliers, we adjust the standardization of features using the median and median absolute deviation (MAD). The formula is given by:

$$Z = \frac{X - \text{median}(X)}{\text{MAD}(X)} \tag{7}$$

where $Z$ represents the standardized feature values, $X$ is the original feature data, median($X$) is the median of each feature, and MAD($X$) is the median absolute deviation of each feature.

*Lasso Regularization with Dynamic Parameter*

We introduce a dynamic regularization parameter in Lasso regression that adapts based on the spread of the coefficients:

$$\lambda = \lambda_0 \cdot e^{-\alpha \cdot \text{std}(\beta)} \tag{8}$$

where $\lambda$ is the adaptive regularization parameter, $\lambda_0$ is the base level of the regularization parameter, $\alpha$ is a scaling factor that influences the rate of exponential decay, and std($\beta$) is the standard deviation of the regression coefficients, $\beta$.

2) *Computation of training data:* Training data for the prediction of Ontology-mapped global attributes is computed. The health data sources use their proprietary attribute names. The automation of the schema mapping process requires training data that can predict the global attribute names of the data sources. The training data is created by performing the statistical computations (min, max, avg, quartiles) on the attributes of each data sources. The statistical computations and their corresponding attribute names are used as the training dataset.

### *3) Machine learning classification*

The machine learning algorithm is applied on the training data that is created in pre-processing phase. The model is trained to predict the global attribute classes.

*Support Vector Machine (SVM)*

$$\text{minimize} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \tag{9}$$

$$\text{subject to} \quad y_i(w \cdot x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0 \tag{10}$$

*K-Nearest Neighbors (KNN)*

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{d}(x_{ik} - x_{jk})^2} \tag{11}$$

*Naïve Bayes*

$$P(y|x_1, \ldots, x_n) = \frac{P(y)P(x_1, \ldots, x_n|y)}{P(x_1, \ldots, x_n)} \tag{12}$$

*Logistic Regression*

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{13}$$

$$z = w \cdot x + b \tag{14}$$

*Neural Network*

$$y = f(w \cdot x + b) \tag{15}$$

*AdaBoost*

$$D_{t+1}(i) = \frac{D_t(i)\exp(-\alpha_t y_i h_t(x_i))}{Z_t} \tag{16}$$

## 6. Experiment Setup

The research work uses the dataset having the observations of one of the vital signs of the human body i.e., Blood Pressure (BP). The datasets are mapped to the proposed Ontology approach using SMOD approach. The mapping accuracy on SMOD is evaluated on various classification algorithms for mapping the data source schemas. The different classification algorithms used for evaluation are: SVM, K-NN, Naive Bayes, Logistic regression, Multi-layer Perceptron (MLP) Neural Networks, and AdaBoost. Naive Bayes, SVM, Logistic Regression, KNN, AdaBoost, and Neural Networks were chosen on the basis of their wide acceptability for handling linear and non-linear classification. Naïve Bayes is based on probabilistic measures of similarity. Its ability to handle large datasets quickly makes it useful for initial alignment steps and for generating probabilistic matches that can be refined further. SVM is applied to ontology alignment by learning a decision boundary that separates matching from non-matching concept pairs based on feature vectors derived from the ontologies. Their ability to handle non-linear relationships using kernel functions is particularly useful in capturing complex similarities between concepts. Logistic Regression probabilistic approach and is helpful where concepts from multiple ontologies found equivalent. KNN can determine alignment based on local patterns in the data, which is helpful when dealing with heterogeneous ontologies with varying structures. AdaBoost can combine different simple classifiers to improve alignment accuracy. It can handle diverse and complex relationships between ontologies. The neural networks capture complex patterns and relationships in data through multiple layers of abstraction. They are highly flexible and capable of learning non-linear mappings.

### *A. Testbed*

The proposed approach is applied to the data sources containing BP observations of a patient. According to the OpenEHR health standard, the BP observation is comprised of four different attributes, these are systolic (sys), diastolic (dys), mean arterial pressure (map) and pulse pressure (pp).

The configuration of the machine is 8 GB of Random Access Memory, 1.80 GHz Core i5 processor and 64-bit Windows 10 Operating system.

The datasets utilized for the attribute mapping approach are gathered from three distinct categories of data sources. These are standardized HL7 Fast Healthcare Interoperability Resources (FHIR) data format (JSON), non-standardized data format (.csv files), and artificial data streams (key-value pair). The datasets store the data in the different data models and have different attribute names to store BP observations. The table II describes different characteristics of datasets used for experimentation such as type, data format and BP Attribute names mentioned in respective datasets such as in Cardio dataset Ap_hi attribute represents for systolic BP and Ap_lo represents diastolic attribute.

Different wrappers are implemented to access the required BP observations from different categories of data sources. There is no dataset available that exclusively stores the BP observations of the patients. Therefore, the desired BP attributes are selected from the available datasets. All the datasets have both Systolic BP and Diastolic BP. However, Mean Arterial Pressure (MAP) and Pulse Pressure (PP) observations was not found in any of the datasets. Therefore, the values of MAP and PP are computed using the standard formula, as shown in equations 17 and 18.

$$MAP = \frac{SYS + 2 * DYS}{3} \tag{17}$$

$$PP = SYS - DYS \tag{18}$$

The details of the dataset used are as follows:

*1) Dataset category-1: Health Standard HL7 FHIR data format*

The HL7 FHIR data format adheres to the JSON data format standard [27]. Only a single artificial record is created in this data format. The wrapper is implemented to fetch the desired BP attributes and their readings. This dataset contains both systolic and diastolic BP attributes.

*2) Dataset category-2: Non-Standard datasets from online repositories*

The non-standard BP observations are taken from two datasets. The first is referred to as ‘FemtoDos’ [28]. It is associated with the prediction relationship between the BP and Body Mass index of patients and another dataset is taken from the cardiovascular disease dataset [29]. The datasets are comprised of 225 and 70,000 records for BP readings respectively. Both of the datasets contain systolic and diastolic BP attributes.

*3) Dataset category-3: Artificially created data streams*

The artificial data stream is created on Kafka producer [30]. It is in key-value pair format. It comprises of total 174 total number of instances. The data stream contains systolic and diastolic attributes.

The overall size of the dataset comprised of 70,400 records of BP observations of different patients.

### B. *Performance evaluation*

The performance of the classification algorithms is evaluated on the basis of the accuracy achieved. The accuracy of a classifier is computed using confusion matrix parameters and k-fold cross-validation accuracy.

### C. *Data-preprocessing*

Data pre-processing involves data cleaning. Data cleaning is the process of transforming and filtering raw data into a usable form. It is performed to remove noisy and incorrect observations from the data set. The dataset contains negative and inaccurate blood pressure readings. The missing values are replaced with the mean value of the attribute and the dataset anomalies are removed using the imputation technique.

The data pre-processing is also performed to generate the training data. It involves computation of statistical measures such as mean, min, max, and quartiles for each attribute in the dataset. These statistical calculations for each attribute are stored as features in individual files, each designated with a target class of SYS, DYS, PP, or MAP. The training data computed requires the statistical features from large number of sources in order to increase the size of training data. Hence, to address this challenge, the dataset from each data source is partitioned into sets of 50 instances. The statistical computations of each group are computed to form training data. Thus, for a dataset of size 70,400 rows, the final training dataset produced has a total of 1,408 instances.

## 7. Experiment Results

The performance of the proposed ontology on SMOD methodology is evaluated on six widely used classification algorithms such as SVM, KNN, Naïve Bayes, Logistic Regression, Neural Network and AdaBoost. The evaluation is conducted based on the various parameters; True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN); shown in table III.

Based on these parameters, we evaluated following performance metrics:

1) Accuracy: It is the simple & most intuitive metric. It measures the proportion of correctly predicted instances out of the total instances.

$$Accuracy = \frac{TP + TN}{Total\ Instances}$$

2) Precision: It measures the accuracy of the positive predictions made by a model. It tells us how many of the predicted positive instances were actually

TABLE II. Experimental dataset for evaluation of SMOD

| **Dataset** | BP[14] | FemtoDos[15] | Cardio[16] | BP Stream |
|---|---|---|---|---|
| **Type** | HL 7 FHIR standard | Non-Standard | Non-Standard | Artificial stream |
| **Format** | JSON | CSV | CSV | Key-value |
| **BP Attributes** | Systolic: systolic BP,<br>Diastolic: diastolic BP | SBP: systolic BP,<br>DBP: diastolic BP | Ap_hi: systolic BP,<br>Ap_lo: diastolic BP | Sys: systolic BP,<br>dys: Diastolic BP |

TABLE III. Model evaluation parameters

| Parameter Name | Description |
|---|---|
| **TP** | No. of correct positive predictions. |
| **FP** | No. of incorrect positive predictions. |
| **FN** | No. of actual positive cases that were incorrectly predicted as negative |
| **TN** | No. of correct negative predictions |

TABLE IV. Comparison of Confusion matrix parameters on different classification algorithms

| Algorithm | SVM | KNN | Naive Bayes | Logistic Regression | Neural Network | AdaBoost |
|---|---|---|---|---|---|---|
| **Accuracy (%)** | 99.6 | 98 | 99.6 | 94 | 47 | 74.5 |
| **Precision (%)** | 99.8 | 98 | 99.7 | 94 | 42 | 64 |
| **Recall (%)** | 99.4 | 96 | 99.5 | 93 | 47 | 75 |
| **F1 Score (%)** | 99.5 | 97 | 99.4 | 94 | 39 | 67 |

positive.

$$Precision = \frac{TP}{TP + FP}$$

3) Recall: It measures how well the model can identify all the positive instances in the dataset. It is also known as sensitivity or true positive rate.

$$Recall = \frac{TP}{TP + FN}$$

4) F1 Score: It is the harmonic mean of Precision & Recall. It provides a single metric that balances both concerns, giving a better sense of the model's performance when you need to consider both false positives and false negatives.

$$F1\ Score = 2 \times \left(\frac{Precision \times Recall}{Precision + Recall}\right)$$

The comparison of Accuracy, Precision, Recall and F1-score of classification algorithms are presented in table IV.

We conducted a comprehensive evaluation of several machine learning algorithms to determine their effectiveness in classification tasks. We used Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naive Bayes, Logistic Regression, Neural Network, and AdaBoost for performance comparison. The performance of these algorithms was assessed based on four key metrics: accuracy, precision, recall, and F1 score.

The SVM demonstrated exceptional performance, achieving an accuracy of 99.6%, precision of 99.8%, recall of 99.4%, and an F1 score of 99.5%. These metrics indicate that the SVM is highly reliable and robust in classifying the data correctly. Similarly, Naive Bayes showed outstanding results with an accuracy of 99.6%, precision of 99.7%, recall of 99.5%, and an F1 score of 99.4%. The KNN algorithm also performed well, with an accuracy of 98%, precision of 98%, recall of 96%, and an F1 score of 97%, though it was slightly less effective in identifying positive instances compared to SVM and Naive Bayes. Logistic Regression also gives the good accuracy of 94%, precision of 94%, recall of 93%, and an F1 score of 94%. This indicates that Logistic Regression is a reliable model but has certain limitations when compared to SVM, Naive Bayes, and KNN. The Neural Network exhibited significantly lower performance, with an accuracy of 47%, precision of 42%, recall of 47%, and an F1 score of 39%. AdaBoost performed moderately, achieving an accuracy of 74.5%, precision of 64%, recall of 75%, and an F1 score of 67%.

The cross-validation is performed to analyze the ability of the algorithm for unseen datasets. It helps to identify the issues such as overfitting. It also provides insights into the generalization of the model. The average accuracy achieved for cross-validation scores for values of k= 5 and k=10 is mentioned in the table V. The results reveal the highest accuracy with a cross-validation score of 10.

The accuracy results of all six classification methods in predicting the global ontology attribute label is compared in figure 4. The experiment results revealed that SVM achieved a CV score of 10, accurately classifying all attribute labels with a accuracy of 99.7%. On the other hand, Naive Bayes and kNN both attain 99.1% accuracy. The logistic regression yields 90.05% accuracy while the multilayer perceptron neural network exhibits 80.5% accuracy. The accuracy for AdaBoost comes out to 81%.

The experimental results of the confusion matrix and cross-validation parameters reveal that both SVM and Naïve Bayes provides the highest accuracy in predicting the correct global class attribute.

## 8. Validation of results

Hypertension is widespread among individuals with cardiovascular, kidney, and diabetic disorders [31]. It elevates the likelihood of both kidney and cardiovascular ailments. It has been seen that patient having diabetes are more likely to have kidney disorders. The abnormal BP is one of the major factors of causing these diseases [32]. Thus, the

TABLE V. Comparison of average accuracy on different cross validation parameters

| Algorithm | Cross Validation | K | Average Accuracy |
|---|---|---|---|
| **SVM** | 5 | - | 99.41 |
| | **10** | **-** | **99.7** |
| **Naive Bayes** | 5 | - | 98.32 |
| | **10** | **-** | **99.1** |
| **Logistic Regression** | 5 | - | 88.61 |
| | **10** | **-** | **90.05** |
| **Neural Network** | 5 | - | 49 |
| | **10** | **-** | **80.5** |
| **AdaBoost** | 5 | - | 79.1 |
| | **10** | **-** | **81** |
| **KNN** | 5 | 3 | 98.81 |
| | 5 | 5 | 98.62 |
| | 5 | 7 | 96.91 |
| | 5 | 9 | 92.83 |
| | **10** | **3** | **99.1** |
| | 10 | 5 | 98.91 |
| | 10 | 7 | 98.33 |
| | 10 | 9 | 98.82 |

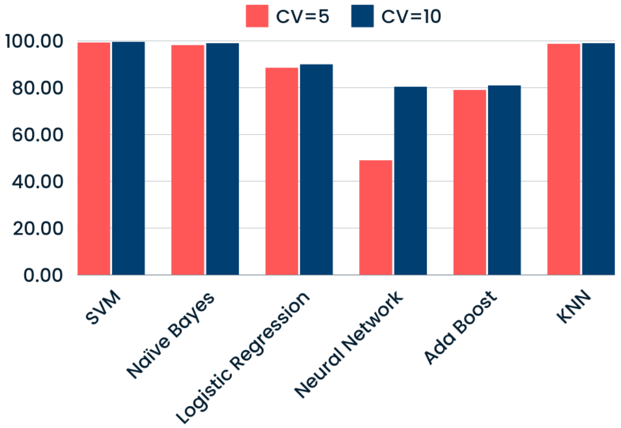

Figure 4. Comparison of accuracy in predicting ontology attribute

datasets of cardiovascular, Diabetes and Kidney disorder may have different ranges of BP values in comparison to other datasets having normal BP readings. Hence, to ensure result generalization and mitigate overfitting, the experimental findings are validated using datasets covering various diseases i.e., Diabetes, Heart, and Kidney.

### *A. Validation datasets*

The experimental results are validated on three different disease datasets. These are Diabetes [33], Heart problem [34], and kidney disease [35]. The details of the validation datasets are presented in table VI. The dataset size consists of 304 heart patients, 768 of diabetes patient records, and 400 of kidney patient records. The desired Blood Pressure attributes are selected from these datasets. To increase the training data size, each attribute is partitioned into a group of 10 data size. The statistical computations are applied to each group and validation is performed for each group of a disease dataset.

TABLE VI. Validation dataset

| Dataset | Instances | BP attributes |
|---|---|---|
| Heart | 304 | trestbps |
| Diabetes | 768 | BloodPressure |
| Kidney | 400 | bp |

### *B. Validation results*

The validation results of the proposed work are shown in table VII. The table shows the misclassification of predicted attributes for different attributes of validation datasets. It shows the percentage of misclassification occurring for different classification algorithms on validation datasets. It comprises of three types of attribute class labels. The local attribute name is the attribute label used in the validation dataset. Actual attribute label is the expected attribute name based on the ontology, and predicted value give the response we get on applying the algorithm.

The validation outcomes shows that most of the algorithms has marginal misclassification rate of less than 2% . It has been revealed that Naive Bayes consistently predicts attribute labels accurately across all three disease datasets with very less misclassification rate, whereas other classification algorithms fail to achieve higher accuracy for at least one of the disease datasets. Although SVM exhibits the highest training accuracy, it fails to accurately identify all three datasets during validation, suggesting potential overfitting of the training data. As illustrated, the logistic regression technique classifies only a single dataset, and AdaBoost classifies only two datasets accurately. With all three disease datasets, the MLP neural network approach shows poor accuracy in predicting the global attributes. It shows the highest rate of misclassification. MLP neural network performs similar to the lower accuracy obtained throughout the training and validation phases.

The validation accuracy score on different disease datasets is shown in figure 5. The results show that Naive Bayes reliably predicts attribute labels across all three disease datasets, whereas other classification algorithms demonstrate lower accuracy in identifying attributes across all the disease datasets.

## 9. Discussion

The experimental results for mapping the proposed ontology demonstrated that the accuracy of SVM and Naïve Bayes classification algorithms is the highest using both confusion matrix and cross-validation. The MLP neural network and AdaBoost showed lower accuracy in both cases. A neural network requires a huge amount of training data to acquire sufficient model accuracy. Therefore, the neural

TABLE VII. Validation Results

| Classifier Name | Dataset Name | Attribute Class Label | | | Misclassified(%) |
|---|---|---|---|---|---|
| | | Local | Actual Value | Predicted Value | |
| **SVM** | **Diabetes** | BloodPressure | DYS | DYS/PP | 29.7 |
| | **Heart** | trestbps | SYS | SYS | 0.2 |
| | **Kidney** | bp | DYS | DYS | 0.4 |
| **KNN** | **Diabetes** | BloodPressure | DYS | DYS/PP | 5.4 |
| | **Heart** | trestbps | SYS | SYS | 1.1 |
| | **Kidney** | bp | DYS | DYS | 1.3 |
| **Naive Bayes** | **Diabetes** | BloodPressure | DYS | DYS | 0.2 |
| | **Heart** | trestbps | SYS | SYS | 0.3 |
| | **Kidney** | bp | DYS | DYS | 0.3 |
| **Logistic regression** | **Diabetes** | BloodPressure | DYS | PP | 54.1 |
| | **Heart** | trestbps | SYS | SYS | 1.1 |
| | **Kidney** | bp | DYS | PP | 35.2 |
| **Neural Network** | **Diabetes** | BloodPressure | DYS | PP | 91.9 |
| | **Heart** | trestbps | SYS | SYS | 76.4 |
| | **Kidney** | bp | DYS | SYS | 81.3 |
| **AdaBoost** | **Diabetes** | BloodPressure | DYS | DYS | 1.6 |
| | **Heart** | trestbps | SYS | DYS | 20.6 |
| | **Kidney** | bp | DYS | DYS | 0.8 |

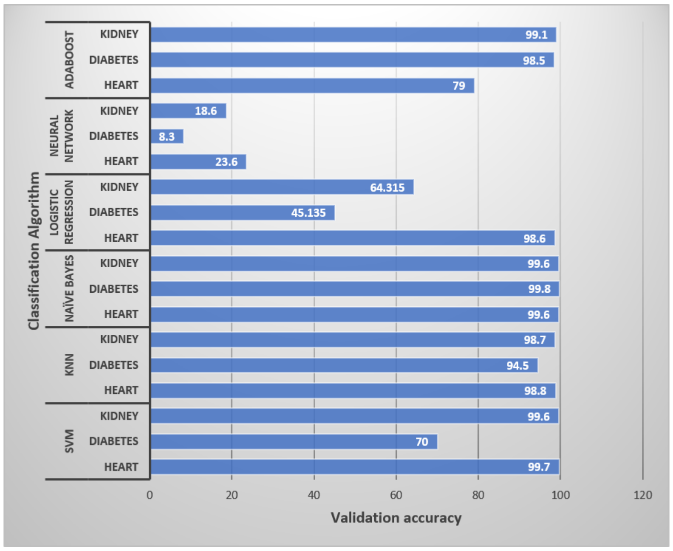

Figure 5. Comparison of validation accuracy on different disease datasets

network shows low classification accuracy in the proposed approach. Similarly, AdaBoost is sensitive to outliers that prevent it from giving high performance.

The validation results indicate that only the Naive Bayes method achieves 99.6% accuracy in predicting attribute labels across all three validation datasets. The reason of it is that the SMOD approach utilizes statistical measures as training features so there is high possibility that the test data BP attribute values lie in their correct ranges. The difference between experimental and validation results of SVM may be that the SVM overfits the training data, whereas the Naive Bayes algorithm does not overfits. Thus, results of the evaluation of the SMOD over proposed Ontology concludes that Naive Bayes shows better accuracy and is considered as one of the promising algorithms among the six for the generalization of the proposed schema mapping.

## 10. Conclusions

The research work carried out proposed the vital signs ontology. The proposed ontology serves as a mediated schema for data integration from diverse health data sources. The query engines use mediated schema for querying different data sources. The schema mapping approach called SMOD map the vital signs of the patients to their corresponding ontology attributes. The ontology-based data integration approach provides global data definitions that enable the access of different data sources in a consistent manner.

The key contribution of the research work carried out includes automation in the ontology alignment process, reducing the need for manual intervention and speeding up the integration process. It also enhances health Data interoperability by enabling seamless data exchange and integration across different health data. The work done provides a scalable and efficient algorithm that can handle large volumes of EHR data, making it feasible for use in large healthcare organizations and research institutions. Integrating patient health data raises several ethical considerations, particularly around data privacy, security, and maintaining patient confidentiality. The proposed approach ensures patient data privacy by computing and exposing statistical properties of patient's health attribute observations, instead of actual health observations.This research would help in improved clinical decision making, public health monitoring, supports personalized medical care and enhanced clinical research. The proposed work is limited to the primary vital signs of the patients. However, it can be further extended to integrate much more medical records such as ECG signals, text and medical images, etc.

## References

[1] A. Doan, A. Halevy, and Z. Ives, *Principles of data integration*. Elsevier, 2012.

[2] H. Liyanage, P. Krause, and S. De Lusignan, "Using ontologies to improve semantic interoperability in health data," *BMJ Health & Care Informatics*, vol. 22, no. 2, 2015.

[3] A.-K. Kock-Schoppenhauer, C. Kamann, H. Ulrich, P. Duhm-Harbeck, and J. Ingenerf, "Linked data applications through ontology based data access in clinical research," in *Informatics for Health: Connected Citizen-Led Wellness and Population Health*. IOS Press, 2017, pp. 131–135.

[4] Clinical knowledge manager. [Online]. Available: https://ckm.openehr.org/ckm/archetypes/1013.1.3574/mindmap

[5] C. Peng and P. Goswami, "Meaningful integration of data from heterogeneous health services and home environment based on ontology," *Sensors*, vol. 19, no. 8, p. 1747, 2019.

[6] A. Kiourtis, A. Mavrogiorgou, A. Menychtas, I. Maglogiannis, and D. Kyriazis, "Structurally mapping healthcare data to hl7 fhir through ontology alignment," *Journal of medical systems*, vol. 43, pp. 1–13, 2019.

[7] A. Cimmino, M. Poveda-Villalón, and R. García-Castro, "ewot: A semantic interoperability approach for heterogeneous iot ecosystems based on the web of things," *Sensors*, vol. 20, no. 3, p. 822, 2020.

[8] N. Gupta and B. Gupta, "Machine learning approach of semantic mapping in polystore health information systems," *International Journal of Computer Information Systems and Industrial Management Applications*, vol. 13, pp. 12–12, 2021.

[9] S. Frid, X. P. Duran, G. B. Cucó, M. Pedrera-Jiménez, P. Serrano-Balazote, A. M. Carrero, R. Lozano-Rubí *et al.*, "An ontology-based approach for consolidating patient data standardized with european norm/international organization for standardization 13606 (en/iso 13606) into joint observational medical outcomes partnership (omop) repositories: description of a methodology," *JMIR Medical Informatics*, vol. 11, no. 1, p. e44547, 2023.

[10] P. Mitra, G. Wiederhold, and J. Jannink, "Semi-automatic integration of knowledge sources," *Proceedings of Fusion'99, July 1999*, 1999.

[11] W.-S. Li and C. Clifton, "Semint: A tool for identifying attribute correspondences in heterogeneous databases using neural networks," *Data & Knowledge Engineering*, vol. 33, no. 1, pp. 49–84, 2000.

[12] J. Madhavan, P. A. Bernstein, and E. Rahm, "Generic schema matching with cupid," in *vldb*, vol. 1, no. 2001, 2001, pp. 49–58.

[13] R. Fagin, L. M. Haas, M. Hernández, R. J. Miller, L. Popa, and Y. Velegrakis, "Clio: Schema mapping creation and data exchange," *Conceptual Modeling: Foundations and Applications: Essays in Honor of John Mylopoulos*, pp. 198–236, 2009.

[14] C. A. Knoblock, P. Szekely, J. L. Ambite, A. Goel, S. Gupta, K. Lerman, M. Muslea, M. Taheriyan, and P. Mallick, "Semi-automatically mapping structured sources into the semantic web," in *Extended semantic web conference*. Springer, 2012, pp. 375–390.

[15] M. Birgersson and G. Hansson, "Data integration using machine learning: Automation of data mapping using machine learning techniques," 2016.

[16] J. Rouces, G. de Melo, and K. Hose, "Complex schema mapping and linking data: Beyond binary predicates." in *LDOW@ WWW*, 2016.

[17] A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun *et al.*, "Scalable and accurate deep learning with electronic health records," *NPJ digital medicine*, vol. 1, no. 1, pp. 1–10, 2018.

[18] J. Xu, X. Xi, J. Chen, V. S. Sheng, J. Ma, and Z. Cui, "A survey of deep learning for electronic health records," *Applied Sciences*, vol. 12, no. 22, p. 11709, 2022.

[19] Clinical knowledge manager. [Online]. Available: https://ckm.openehr.org/ckm/archetypes/1013.1.3574

[20] Clinical knowledge manager. [Online]. Available: https://ckm.openehr.org/ckm/archetypes/1013.1.2796

[21] Clinical knowledge manager. [Online]. Available: https://ckm.openehr.org/ckm/archetypes/1013.1.4295

[22] Clinical knowledge manager. [Online]. Available: https://ckm.openehr.org/ckm/archetypes/1013.1.4218

[23] R. Mizoguchi, E. Sunagawa, K. Kozaki, and Y. Kitamura, "The model of roles within an ontology development tool: Hozo," *Applied Ontology*, vol. 2, no. 2, pp. 159–179, 2007.

[24] A. Kalyanpur, B. Parsia, E. Sirin, B. C. Grau, and J. Hendler, "Swoop: A web ontology editing browser," *Journal of Web Semantics*, vol. 4, no. 2, pp. 144–153, 2006.

[25] V. Jain and M. Singh, "Ontology development and query retrieval using protégé tool," *International Journal of Intelligent Systems and Applications*, vol. 9, no. 9, pp. 67–75, 2013.

[26] Protégé. [Online]. Available: https://protege.stanford.edu/

[27] Example observation/blood-pressure (json). [Online]. Available: https://www.hl7.org/FHIR/observation-example-bloodpressure.json.html

[28] H. Golino, "Women's dataset from the Predicting increased blood pressure using Machine Learning paper," 11 2013. [Online]. Available: https://figshare.com/articles/dataset/Women_s_dataset_from_the_Predicting_increased_blood_pressure_using_Machine_Learning_paper/845664

[29] Cardiovascular disease dataset. [Online]. Available: https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset

[30] Apache kafka. [Online]. Available: https://kafka.apache.org/intro

[31] P. N. Van Buren and R. Toto, "Hypertension in diabetic nephropathy: epidemiology, mechanisms, and management," *Advances in chronic kidney disease*, vol. 18, no. 1, pp. 28–41, 2011.

[32] H. Murtaza, M. Iqbal, Q. Abbasi, S. Hussain, H. Xing, and M. Im-

ran, “Correlation analysis of vital signs to monitor disease risks in ubiquitous healthcare system,” *EAI Endorsed Transactions on Industrial Networks and Intelligent Systems*, vol. 7, no. 24, 2020.

[33] Pima indians diabetes database. [Online]. Available: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

[34] S. W. P. M. Janosi, Andras and R. Detrano, “Heart Disease,” UCI Machine Learning Repository, 1988, DOI: https://doi.org/10.24432/C52P4X.

[35] S. P. Rubini, L. and P. Eswaran, “Chronic Kidney Disease,” UCI Machine Learning Repository, 2015, DOI: https://doi.org/10.24432/C5G020.