International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 13 (2021) pp. 222-232

Received: 18 September, 2020; Accepted: 10 June, 2021; Published: 12 August, 2021

Machine Learning Approach of Semantic Mapping in Polystore Health Information Systems

Nidhi Gupta[1] and Bharat Gupta[2]

[1] Department of Computer Science and Engineering, Jaypee Institute of Information Technology,
A 10, Industrial Area, Sector 62, Noida, Uttar Pradesh 201309
nidhi.gupta.it@ipec.org.in

[2] Department of Computer Science and Information Technology, Jaypee Institute of Information Technology,
A 10, Industrial Area, Sector 62, Noida, Uttar Pradesh 201309
bharat.gupta@jiit.ac.in

***Abstract*: Health analysis and Information system use patient data from ubiquitous data sources for decision making. Limited adoption of health standards results in difficult data exchange and its reuse. The data integration from Polystore databases experience schema level conflict, which limits its sharing and reuse among other health organizations. The research work proposed an approach called Semantic Mapping of Observation Data for interoperability of Polystore Electronic Health record data sources. The proposed approach resolves schema level conflicts that occur while integrating patient EHR from multiple heterogenous data sources. The research work demonstrates SMOD on Blood Pressure data from standardized, non-standardized and streaming data sources. Its performance is accessed on widely used multiclass algorithms such as Support Vector Machine, K-Nearest Neighbor, Naive Bayes, Logistic Regression and Neural network. Results shows highest accuracy with Linear SVM in comparison with other classification algorithms. However, K-Nearest Neighbor and Naive Bayes performance is nearly close to SVM. The result is validated on Blood Pressure data taken from datasets of 3 different diseases such as Heart, Kidney and Diabetes. The validation results demonstrate that Naïve Bayes algorithm is used as best generalized algorithm in SMOD and is able to predict accurate mapping with other Blood Pressure datasets of different diseases.**

***Keywords*:** Machine Learning, Polystore, Semantic Mapping, Integration, Blood Pressure**.**

I. Introduction

In the era of development of modern tools and advanced communication technology, high volume of patient health data is generated and stored in various data formats such as CSV, JSON, RDB etc. The variety of data stored at distributed data sources is referred to as 'Polystore' databases. Health standards are developed to ensure uniformity in storing, accessing, communicating the Electronic Health Record (EHR) such as HL7, OpenEHR, supporting standard vocabulary such as SNOMED, LOINC, ICD, etc. [1]. Insufficient adoption of standardized methods of data storage and retrieval results in limited sharing of EHR and prolong delay in patient monitoring. Health systems uses various methods and technologies for monitoring patient health and requires integration of static data and dynamic data streams for efficient clinical diagnosis and health decision making[2]. Some companies such as Amazon, Google have also initiated to provide the models for predicting heath data by maintaining patient EHR. Thus, there is a need of data integration from autonomous data sources in distributed heterogeneous environment.

In healthcare, many applications such as disease diagnosis, remote health monitoring, requires to integrate data from multiple health providers. The patient Electronic Health Records (EHR) is distributed at multiple health providers. Each database stores the data with its own local attribute names. Thus, integration of such data from multiple databases results in naming conflicts. The research carried out proposed a machine learning model called Semantic Mapping of Observation Data (SMOD) to resolve schema level naming conflicts occur during integration of EHR from Polystore databases. The research work carried out enables semantic integration of Patient's Blood Pressure (BP) records from different data sources.

The work is organized into various sections. Section 2 presents the background and motivating scenario of our research problem. Section 3 reviews the related work on semantic mapping for data integration. Section 4 discuss about widely used multiclass classifiers. Section 5 describes the proposed approach used for semantic mapping. Section 6 shows the experimental evaluation of proposed approach and validation of its results. Section 7 concludes the research work.

II. Background and Motivation

Data mapping facilitates retrieval, integration and interoperability. This section presents the significance of data mapping in data integration and also presents the motivation scenario behind the integration. Data integration is the process of combining the data from various data sources to provide a single uniform data view. With the increase in volume and

variety of data, manual data integration is not possible. Many data integration and transformation tools are developed to automate the process of data integration. It automates various tasks for better understanding of semantics of data such as entity resolution, data fusion, data extraction. Entity resolution refers to the identification of all existence of an entity across multiple data sources. Data fusion refers to merging of records by resolving data level conflicts from multiple sources to make the data more useful, Data extraction refers to the mining of structured data from semi-structured data sources such as web pages, social media etc.

Data mapping deals with matching of attributes belongs to the same entity in multiple databases. It is required when organizations do not store the data in a standardized format. Data integration process has also utilized various supervised and unsupervised machine learning algorithms such as neural networks, logistic regression etc. for implementing data integration tasks. Machine learning algorithms automates the process of data mapping and gives precise results.

The motivation behind our clinical data integration is to provide usability and accessibility of patient's clinical history of BP measurements, from multiple health providers. Physiological health measures such as body temperature, heart rate, Blood Pressure (BP), etc. are the integral part of any diagnosis. High blood pressure is one of the major causes of Hypertension and Cardiovascular problems among patients. It is a root cause of many deadly diseases and may results in a life-threatening body state such as heart attack, stroke, disorders of pregnancy, cognitive decline, thyroid and chronic kidney disease. Nearly, half of the adults in United States have hypertension, defined as systolic BP greater or equal to 130 mm Hg or a diastolic BP greater than equal to 80 mm Hg[3]. In India, hypertension cases are increasing at alarming rate and every 1 out of 3 people is suffering from it. The symptom of Hypertension is not explicitly visible and is considered as a silent killer. Ambulatory Blood Pressure Monitoring (ABPM) monitors the individual BP during routine work at regular intervals by wearable sensor device. ABPM measurements are analyzed when patient feels anxiety and nervousness in clinical settings. Therefore, Blood Pressure needs to be properly measured and analyzed for taking any health decision and is considered as one of the prime physiological factors. BP analysis of a patient requires integration of clinical measurement stored with various health providers.

The sequence of steps for query processing in Polystore data environment is shown in figure 1. It has three layers i.e. User Interface, Middleware and Data storage. The user query for retrieving and integrating the data from distributed data formats use a common query interface. The middleware consists of query processing engine, which scans the query and search for the relevant data sources to execute the query. The query gets executed at respective data sources and results produced gets merged in response to user query. Each clinical data source, stores the local view of its data. A unified data view requires to map the local data view to its corresponding global view. In the research carried out, data view (local and global) refers to the attribute labels used to describe the data. The input query at user interface, use standardized global name to refer each attribute of required data. To understand the query attributes at each autonomous data source, it is essential to map the global attributes to their corresponding local attribute.

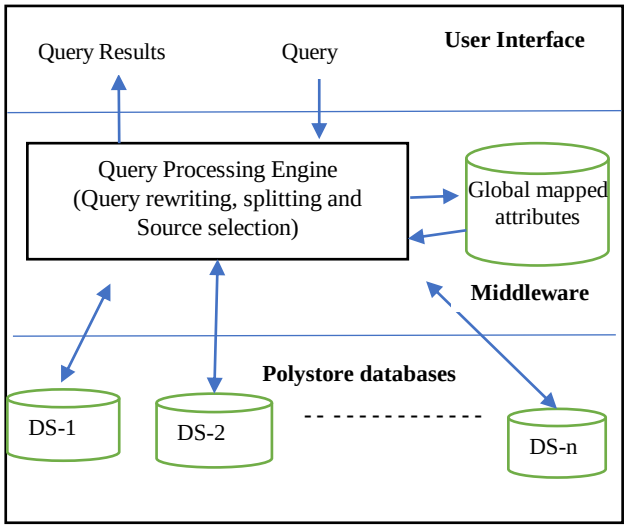

Figure 1. High level view of Query Processing in Polystore Databases

III. Related work

Semantic mapping and integration are essential to an interoperable system. Much of the work has been done on data mapping for enterprise database integration. E. Rahm et al. [4] and B. Gu et al. [5] worked on database matching levels such as schema, record, element, instance and structure levels.

M. Birgersson et al. [6] automates data mapping of different XML documents by comparing xPath of different data source. It uses network data flow model to find different words with same semantic meaning.

W.S Li et al. [7] proposed a technique for semantic matching of equivalent data elements. It uses the metadata and trains a neural network for data integration.

The existing literature also provides solutions for schema and concept mapping in health data [8]. M. Krol et al. [9] studied the issues of integrating multi-platform medical systems and discuss its possible solutions.

A. Roehrs et al. [10] introduced a methodology named OmniPHR (Personal health record) for semantic interoperability and integration of different health standard using Natural Language Processing and Machine learning.

Y. Yang et al. [11] presented a hybrid cloud called Medshare for health professionals to access and share medical information securely. X. Yang et al. [12] proposed a cost-effective integration solution of healthcare system for interoperability in distributed health systems.

A. Brugues et al. [13] focused on achieving interoperability through HL7 CDA standard in health systems for managing gestational diabetes mellitus. N. Gupta et al. [14] discussed about interoperability issues and surveyed about various health standard methods. Kumar et al. provides a framework for generic Information exchange using HL7 in heterogeneous, autonomous systems for data exchange and interoperability [15].

The researchers also worked on mapping of data streams for

efficiently processing, analyzing, joining and querying streaming data. [16] studied the problem of similarity join of data streams. It proposed a model based on updating distance metric of data stream sliding window. S. Benny et al. [17] worked on the problem of entity resolution in streaming data. The similarity score for entity mapping is generated using Hadoop MapReduce framework and is used for matching data streams.

Erhard Rahm et al. [18] compared various existing approaches for semantic mapping. Table 1 shows the characteristics of proposed model SMOD with other state-of-the-art semantic mapping approaches such as SEMINT [19], SKAT [20], DIKE [21], Cupid [22] and Clio [23]. SEMINT is a tool used for semantic integration of heterogenous database. It trains neural network for attribute identification and matching. SKAT is an articulation tool designed for semantic integration of data sources. DIKE is the tool for integration of federated data sources. It proposed an algorithm to form a global dictionary for all possible synonyms for a unified view of data sources. CUPID is an algorithm designed for schema mapping, independent of particular data model. It uses linguistic and structure mapping approach for data integration. Clio also worked for finding the relationship between data in heterogenous data sources for data integration and its exchange. It generates the queries for finding the data mappings in schemas of two heterogenous data sources.

The proposed research work, provides data mapping for data interoperability and its integration with various Health Information System (HIS) stored at Polystore databases. The work carried out comprises of semantic mapping of data from three different category of data sources i.e., data in standardized health data format such as HL7, non-standardized data sources (local databases) and streaming data. The data attributes collected is according to the suggestions of OpenEHR archetype [24].

	SEMINT [19]	**SKAT [20]**	**DIKE [21]**	**Cupid [22]**	**Clio [23]**	**SMOD (Proposed)**
Schema types	relational	XML, IDL	Entity-Relation (ER)	XML, Relational	XML, Relational	CSV, JSON, Key-value pair (data stream)
Match granularity	Element level: attributes	Element/Structure level: attributes/class	Element/Structure level: entity/ relationships/ attributes	Structure level	Element/Structure level: attributes/class	Element/Structure level: attributes
Match cardinality	1:1	1:1, n:1	1:1	1:1, n:1	1:1, n:1, n:m	1:1 (Data in accordance to OpenEHR archetype)
Preprocessing required	Yes	Yes	Yes	Yes	Yes	Yes
Technique used	Clustering using Neural Network	Rule based approach in first order logic	Determine relationship between object of different entity(is-a)	Linguistic and structural similarity	Use semantic mapping queries for finding the relationship between two schemas.	Use multiclass classifiers. Best result with Naïve Bayes

Table 1. Characteristics of proposed model- SMOD with state-of-the-art schema mapping approaches

The contribution of the paper is as follows:

1. The research work proposed an automated machine learning approach to resolve the semantic mapping problem. It focuses on resolution of attribute naming conflicts for integration of distinct databases.
2. The pre-processing step of proposed approach generates training dataset with reduced noise. The proposed approach computes statistical measures for computing features of training data during pre-processing step. Therefore, the statistical summarization reduces or eliminate the effect of noise in the data.
3. The work done facilitates mapping on blood pressure attributes taken from static and data streams data sources.

IV. Multiclass Classification Algorithms

A. Support Vector Machine

SVM is one of the widely used classification algorithm. It works by plotting the training data in n-dimensional space and separates each class using a decision boundary called hyperplane. In n-dimensional space, hyperplane is of n-1 dimension i.e., in 2-D the hyperplane is a line, in 3-D it is a plane. Therefore, dimensions in hyperplane depends on the number of features in training data. Support vectors are the

data points which are at minimum distance from hyperplane. There are many possible hyperplanes to separate the training data. The algorithm strives to find the optimal hyperplane, where the distance between the support vectors and hyperplane should be maximum. SVM are well known for classification of high dimensional data i.e., dataset with large number of features than its instances. The algorithm can be used to classify both linear and non-linear data with the help of mathematical function called 'Kernel'. A kernel trick transforms the input data into high dimension space. It helps to find an optimal hyperplane to separate data points of each target classes. The research work carried out uses kernel function as 'linear' for the classification problem. The generalized representation of kernel function is (1)

$$k(x,y) = \big(\varphi(x), \varphi(y)\big) \qquad (1)$$

where, x and y are input data s.t.

$$x, y \in X$$

φ is a mapping function for mapping input data to high dimension given in

$$\varphi: X \to R^n \qquad (2)$$

The value of kernel function $k(x,y)$ returns the similarity score of input data points in n-dimension space.

SVM is primarily used for classification on binary classes, but it is also used for multiclass classification using one vs. all approach. Using this approach, if c is the number of target classes, then c binary SVM models are constructed, one for each class. In general, the new sample x_i belongs to those class C_j, with largest decision function(y′) value given as in equation (3)

$$y' = \arg max \quad w_j{}^T x_i + b_j \qquad (3)$$

If $y' \le -1, then\ y' \ne C_j$
$y' \ge 1, then\ y' = C_j$

For a non linear SVM, the decision function f(x) is given as in equation (4)

$$f(x) = sgn\left(\sum_i^N y_i \alpha_i K\big(x_i, y_j\big) + b\right) \qquad (4)$$

where,
w= coefficient vectors
N= number of training points
α_i= coefficient of Lagrange multipliers

B. K-Nearest Neighbor

In KNN, the class prediction is highly influenced by the most common class of k-nearest training data points. The algorithm gives best results with low dimensional data i.e., dataset with a smaller number of features than instances. For optimal execution of KNN classification, the research work done found the best value of K by comparing the accuracy achieved with distinct k values.

C. Naïve Bayes

Naïve Bayes classification is based on the Bayes theorem with the assumption of independence of features. Each feature in the dataset is considered as independent with one another, and hence is termed as 'Naïve'. It uses the probability to make predictions on input data instance to the target class. The Bayes theorem is stated as shown in (5)

$$P\left(\frac{Y}{X}\right) = \frac{P\left(\frac{X}{Y}\right) * P(Y)}{P(X)} \qquad (5)$$

where,

P(Y)=Probability of a class
P(X)=Probability of a data
P(X/Y) = Probability of data, given the class is true
P(Y/X) = probability of a class, given the data

The research work carried out uses Gaussian Naïve Bayes function. It is used when x_i is a real value and $P(x_i/y_j)$ exhibit a normal distribution of data as shown in (6)

$$P\left(\frac{x_i}{y_j}\right) = \left(\frac{1}{\sqrt{2\pi\sigma_j{}^2}}\right) \exp\left(-\left(\left(x_i - \mu_j\right)^2 / (2\sigma_j{}^2)\right)\right) \qquad (6)$$

where,
x_i is the input data and y_j is its specified class.
μ and σ^2 are the mean and variance of each specified class.

D. Logistic Regression

Logistic Regression is used for predicting the data with discrete target classes. It generally gives best results with binary classification problems i.e with two classes. However, it can be used for multiclass classification problems by using 'one vs. all' approach. With this approach, the classifier solves a multiclass problem by reducing it to binary classification problem. The classifier is trained for 'one' class and treat rest classes as 'other'. Then the same it does the same thing with other classes. For a given input, all the trained classifier gets executed and the most confident classifier would be selected as target class. Logistic regression uses the sigmoid function which ranges between 0 and 1. The sigmoid function map the predicted value to the probabilities in range of 0 and 1 only. The (7) shows the sigmoid function and the predicted output is calculated as shown in (8)

$$y = b_0 + b_1 * x \qquad (7)$$

$$P = \frac{1}{1 + e^{-y}} \qquad (8)$$

where

b_o, b_1= input coefficients
P= Predicted output

E. Neural Networks

The basic element of neural networks is called neurons. It has one input layer, one output layer and multiple hidden layers. The layers are composed of predefined number of neurons and connected with neurons of other layers. Each layer receives an input, performs summation of weights and input data, and by. The networks trained iteratively and with each iteration, weights assigned to each neuron are adjusted to reduce square error. Neural network is best used in applications which require learn the complex functions. It is one of the best classifiers for high dimensional problems, but requires a large training data samples to build the classifier. The proposed work uses Multi-layer Perceptron Algorithm (MLP) for construction of classifier. It trains dataset to learns a function, g() given in (5)

$$g: R^m \rightarrow R^n \qquad (9)$$

where,
m = number of input dimensions
n = number of output dimensions

V. Proposed Approach: Semantic Mapping of Observation Data

This section presents a supervised machine learning approach used for data mapping of observation data from numerous data sources. The proposed approach first computes the training data. Thereafter, it implements the data mapping using classification algorithms.

A. Training Data Generation Technique

A machine learning classification algorithm requires training data to classify each local labelled attribute to its corresponding global attribute. The training data is generated before the execution of the proposed approach-SMOD. Training data is formed by the collection of observational datasets from various data sources. Statistical properties (St) like mean, min, max, quartiles for each attribute (A_i) data values, is computed separately, and the attribute labels are manually mapped to global attributes. The computed properties and its corresponding attribute name(global) are maintained in a separate dataset. The dataset formed is used for training data with computed statistical properties as features and its corresponding attribute label as target class.

Algorithm 1. Generation of Trained model

---

1. **Input**: DS= {S_1, S_2…S_m} set of all data sources
   A= {A_1,A_2,…A_k} is the set of all attributes in DS
2. **Output**: Trained classifier : TrainedModel()
3. **for** all data source S_i in DS **do**
4. | **for** each attribute A_j in A **do**
5. | | St=ComputeStatistics(A_j)
6. | | St=AddTargetLabel(A_j, St)
7. | | T_A=AppendRow(T_A, St)
8. | **end**
9. **end**
10. TrainedModel() =Classifier(T_A)

The Algorithm 1 demonstrate the generation of training data (T_A) from various data sources (DS) such that, DS= {S_1, S_2…S_m}and A={A_1,A_2,…A_k} is the set of all attributes in a DS. The steps 3 to 9 in the algorithm generates training data and finally the model is trained in step 10. ComputeStatistics() function in step 5 is used to compute the statistical properties of each attribute as training feature. These properties are minimum value, maximum value, mean value and quartiles of attribute data values. In step 6, AddTargetLabel() assigns the global attribute label of corresponding attribute as target class. The steps 4 to 6 constitute a row in training dataset and in step 7, AppendRow() adds the row in training dataset (T_A). After computation of training data, the machine learning model (TrainedModel()) is trained using classification algorithm(Classifier) in step 10. The trained model predicts the attribute labels of new unseen statistical features.

B. Data Mapping

The main goal of the approach is to perform semantic mapping of input dataset. The architecture of SMOD model is shown in Figure 2.

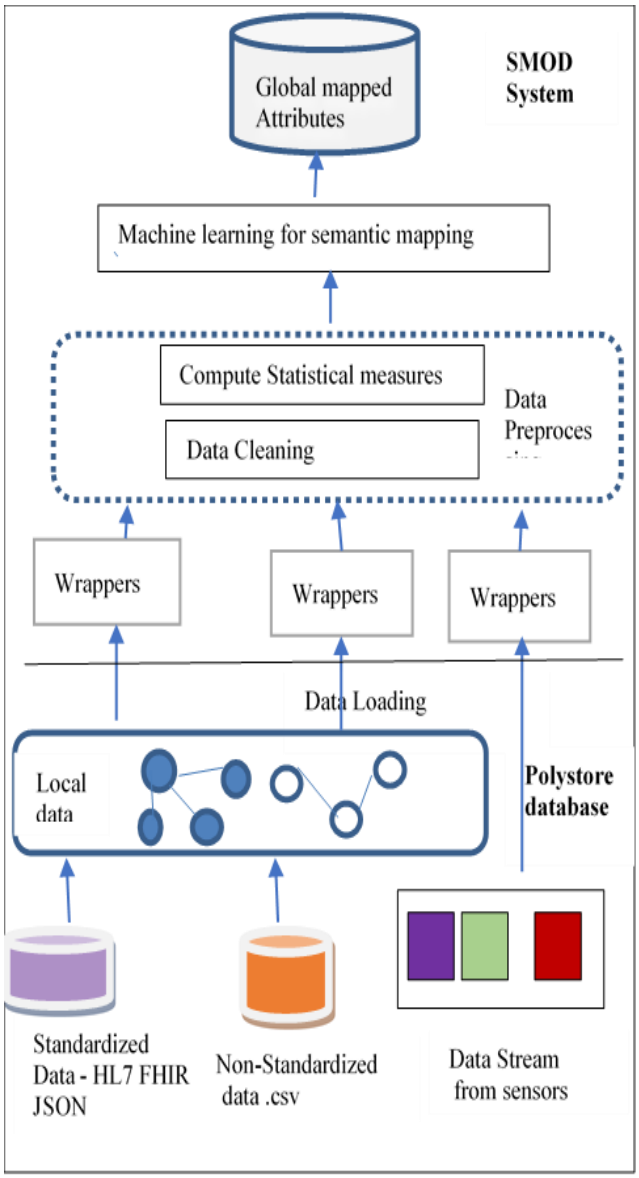

Figure 2. Architecture of SMOD System.

It comprises of three phases:

1) Data extraction: It collects the dataset from data sources. The data set used in the approach are from both static and streaming data. Various wrappers are implemented to extract the multi-format data.

2) Preprocessing: Preprocessing is an integral step to prepare the data to perform mapping. This step performs two tasks i.e. data cleaning and computation. Data cleaning improves the quality of data. It is required as real data often contains missing and incorrect values. Subsequently, statistical measures such as mean, max, min, quartiles are calculated for all attributes in the dataset of in the dataset.

3) Machine learning for semantic mapping: A supervised machine learning algorithm is implemented for mapping local attribute label to its global attribute label. The proposed approach, use a multi-class classification algorithm on computed statistical measures, as computed in previous phase.

The local and global mapped attributes are stored in a database. It is used in various applications that requires data retrieval and integration from Polystore databases, as discussed in section 2.

Algorithm 2. Attribute Mapping

1. **Input:** T_A : Training data
   A[ ] : dataset of attribute required to map
2. **Output:** Global mapped attribute label : Map_A
3. TSt= ComputeStatistics(A[ ])
4. Map_A= TrainedModel(TSt)
5. **return** Map_A

The Algorithm 2 shows the steps performed for attribute mapping. A[] is the input test dataset values of attribute, A. The algorithm first computes statistical measures on attribute values using ComputeStatistics(), thereafter, the training model (TrainedModel()) accepts the statistical values as input features and returns the global attribute mapping(Map_A).

C. SMOD Complexity

The complexity of the proposed approach SMOD, lies in the generation of training datasets as shown in Algorithm 1. To generate n number of training data, n number of distinct datasets are required. The statistical attributes are computed for each attribute of distinct data sources. In generation of class balance training dataset, if k is the number of attributes in each dataset, then statistical attributes are computed for all (n*k) attributes.

Calculation of statistics for an attribute is a constant time operation. Therefore, the worst-case complexity for training set generation is O(n*k).

To apply the proposed approach for mapping of attribute A (Algorithm 2), the statistical properties are computed and trained model is applied on it to produce the mapped global output. Thus, complexity of applying proposed approach is proportional to the time taken by chosen classifier. Therefore, applying the model does not cause any additional overhead except complexity of classifier.

VI. Experimental Evaluation and Results

This section describes the experimental setup and discuss th evaluation results, on applying multiclass classification algorithms in the proposed approach. The validation of results is also performed on different datasets to find the best suitable algorithm.

A. Datasets

The patient health is monitored during each visit at hospitals as well as with the use of health monitoring wearable devices. The limited use of health standards results in data storage in standardized and non-standardized data formats. In view of it, the non-invasive blood pressure readings are taken from three categories of data sources. The first dataset is synthetically created BP observations in standardized HL7 FHIR in JSON format [25]. The second category is non-standardized dataset taken from online repositories. The third data source is streaming BP data created synthetically. There are various factors affecting BP of an individual such as age, height, sex. In addition to this, people suffering from disease shows difference in their BP observations Therefore, non-standardized category of BP observations is selected from two datasets; one associated with predicting relation between BP and Body Mass Index [26] and other from cardiovascular disease dataset in CSV file format [27]. The dataset contains 70,400 instances of different patients. Since, there is no datasets available which only contains BP attributes, so the desired attributes are selected from different data sources using various software wrappers. The experiments are carried out on localhost network. The research work makes the assumption that the attribute used for BP data are in accordance with one as suggested by OpenEHR standard.

B. Data Preprocessing

Preprocessing is performed on the relevant BP features such as Systolic BP(SYS), Diastolic BP(DYS), Pulse Pressure (PP), Mean Arterial Pressure (MAP). Data cleaning transforms and filters the raw data into useful format. Datasets taken from various data sources contain negative BP values, incorrect BP readings (in thousands). Such anomalies are removed using imputation technique by replacing impurities from their attribute mean value.

In case of streaming data, the artificial data streams are generated on Kafka producer in Key-value pair format. The statistical measures are computed on Key-value data stream. The resulted computations are stored in CSV data file format. The computed datastore is then trained just like other static data sources for predicting the attribute global class labels for new datasets. However, all the clinical datasets do not record all the BP attributes readings as suggested by OpenEHR archetype. Therefore, the remaining features values is computed from the given systolic (SYS) and Diastolic (DYS) data, such as Pulse Pressure (PP) and Mean Arterial Pressure

(MAP) using the standard formula as:

$$PP = SYS - DYS \quad (10)$$

$$MAP = (SYS + 2 * DYS)/3 \quad (11)$$

Statistical measures such as mean, min and max are computed for each attribute of datasets. The result of statistical computations of each attribute are stored in a separate file as features with its known target class as SYS, DYS, PP or MAP.

In the proposed approach, the statistical properties of one attribute from a data source contributes for building one instance of training data. Collecting data from such large number of data sources in not possible, so we divide the dataset collected in group of 50 instances and computed the statistical properties of each group. This technique results in 1408 instances of training data.

C. Training and Testing SMOD model

The training dataset has 4 target classes SYS, DYS, PP and MAP. The following points are considered in selection of multi-class algorithms. These are:

a) Since, the boundaries of normal BP attributes are clearly defined [28], it is less than 120 for systolic, less than 80 for diastolic, between 40 and 60 for pulse pressure, and between 70 and 100 for mean arterial pressure. It indicates high potential for linear separable decision boundaries [29].

b) Some raised Blood Pressure data values i.e., systolic blood pressure is greater or equal to 120 or diastolic value is greater or equal to 80, shows class overlapping with normal range of BP values. However, computation of statistical measures of such datasets normalizes the data values and thus reduces the overlapping problem.

c) SVM, KNN, Naive Bayes and Neural network are the multi-class classification algorithms which are used for statistical inference and machine learning[30]. They also worked well to capture meanings and semantic relationships in textual data[31], for instance, Naïve bayes and KNN are used in [32] for schema mapping while Neural network classification performed well in [7] for mapping various data sources.

The proposed approach is compared on 5 different multiclass classifiers i.e., Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Naive Bayes, Logistic Regression and Neural network algorithms. In general, Naïve bayes, and logistic regression algorithms are used for linear separable datasets, while SVM and KNN both works for linear and non-linear classification problems[33]. Neural network ability to train non-linear and complex relationship in data.

The classification algorithms are evaluated using two techniques:

1) k-fold Cross Validation (CV)

In cross validation technique the given data sample is divided into k subsets, out of which one subset act as a test data and rest all is training data. The classifier is trained with training data and evaluate with test data. The process repeats k times with different subset of test data. The most common value of parameter k is 5 and 10. The Tables 2 to 6 depicts the average accuracy achieved in each fold, corresponding to each cv values. The results demonstrate that the classification algorithms show higher accuracy when evaluating it using 10-fold cross validation.

CV	Avg. Accuracy
5	99.4
10	**99.7**

Table 2. SVM Cross validation score

K	CV	Avg. Accuracy
3	5	98.8
5	5	98.6
7	5	96.9
9	5	92.8
3	**10**	**99.1**
5	10	98.9
7	10	98.3
9	10	98.8

Table 3: KNN Cross validation score

CV	Avg. Accuracy
5	98.3
10	**99.1**

Table 4. Naïve Bayes Cross validation score

CV	Avg. Accuracy
5	88.6
10	**90.05**

Table 5. Logistic regression Cross validation score

CV	Avg. Accuracy
5	49.0
10	**80.5**

Table 6. Neural Network Cross validation score

Figure 3. Comparison of cross validation average accuracy results.

The Figure 3 compares the accuracy of all 5 classification algorithms. The test results revealed that for cv score of 10, SVM classified all the attribute labels with 99.7% accuracy. However, KNN and Gaussian Na ïve Bayes both achieves the same accuracy of 99.1%. Logistic regression shows 90.05% and Multilayer perceptron neural network shows 80.5% accuracy in predicting global attribute mapping.

2) Confusion matrix

In addition to cross validation score, the classifiers are also compared on the basis of the different parameters of confusion matrix. It summarized the performance of classification algorithms on test data. The table 7 defines the confusion matrix for multiclass classifier.

		Predicted class			
		SYS	**DYS**	**PP**	**MAP**
Actual Class	**SYS**	T_{SS}	T_{SD}	T_{SP}	T_{SM}
	DYS	T_{DS}	T_{DD}	T_{DP}	T_{DM}
	PP	T_{PS}	T_{PD}	T_{PP}	T_{PM}
	MAP	T_{MS}	T_{MD}	T_{MP}	T_{MM}

Table 7. Confusion Matrix for multiclass classifier

* Here in the table, S, D, M, P notations are used to represent classes SYS(S), DYS(D), PP(P) and M(MAP).

In table 7, the notation $T_{c1,c2}$ represents the number of predictions when actual class is c1 and predicted class is c2. The basic terms used to represent Confusion matrix are:

a) True Positives (TP): In each value of confusion matrix $T_{c1,c2}$ if c1 and c2 both are same then it signifies the correct prediction of that class. For e.g. TP for class SYS is T_{SS} .

b) True Negatives (TN): For a class c1, its TN is the sum values in all row and column of all classes except for c1. For e.g., TN of class SYS is calculated as (12)

$$TN = T_{DD} + T_{DP} + T_{DM} + T_{PD} + T_{PP} + T_{PM} + T_{MD} + T_{MP} + T_{MM} \quad (12)$$

c) False Positives (FP): For a class c1, its FP is the sum of all values in column of class c1, except its TP value. For e.g., FP of class SYS is calculated as

$$FP = T_{DS} + T_{PS} + T_{MS} \quad (13)$$

d) False Negatives (FN): For a class c1, its FN is the sum of all values in row of class c1, except its TP value. For e.g., FN of class SYS is calculated as (14)

$$FN = T_{SD} + T_{SP} + T_{SM} \quad (14)$$

The research work also compares the algorithm with various performance metrics, calculated using Confusion matrix. These are:

a) Precision

For a class c1, Precision is calculated as (15)

$$Precision_{c1} = \frac{TP_{c1}}{TP_{c1} + FP_{c1}} \quad (15)$$

For multiclass classifier, the macro average of Precision is calculated as (16)

$$Precision = \frac{\sum_{i=1}^{N} Precision_{ci}}{N} \quad (16)$$

b) Recall

For a class c1, Recall is calculated as

$$Recall_{c1} = \frac{TP_{c1}}{TP_{c1} + FN_{c1}} \quad (17)$$

For multiclass classifier, the macro average of Precision is calculated as:

$$Recall = \frac{\sum_{i=1}^{N} Recall_{ci}}{N} \quad (18)$$

(c) F1-score

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (19)$$

(d) Accuracy

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (20)$$

The Table 8 represents the comparison of different multiclass algorithms on basis of Precision, Recall, F1-score and Accuracy

Algorithm	**Precision**	**Recall**	**f1-score**	**Accuracy(%)**
SVM	1	1	1	**100**

KNN	.98	.96	.97	98
Naïve Bayes	1	1	1	**100**
Logistic Regression	.94	.93	.94	94
Neural Network	.42	.47	.39	47

Table 8. Comparison of multiclass algorithms on different confusion matrix parameters.

D. Validating the SMOD model

The validation is performed separately on different dataset as hypertension is common among the patients with cardiovascular, Kidney and diabetes diseases[34]. The research work in [34] provides the study about the way hypertension increases the risk for kidney and cardiovascular disease. It also explains that diabetes is most common cause of kidney disease. The dataset of these diseases contains patients suffering from hypertension, and may have different data distribution and different statistical values such as mean value of systolic and diastolic [35]. The promising results of validating the model on these datasets, confirms the proposed approach accuracy on other BP datasets with different data distributions. Validation also prevents overfitting and generalizes of results.

		Attribute class labels		
Dataset	**Local**	**Predicted**	**Actual**	**Misclassified(%)**
Algorithm: SVM				
Heart	trestbps	SYS	SYS	0
Diabetes	BloodPressure	DYS/PP	DYS	29.7
Kidney	bp	DYS	DYS	0
Algorithm: KNN				
Heart	trestbps	SYS	SYS	0
Diabetes	BloodPressure	DYS/PP	DYS	5.4
Kidney	bp	DYS	DYS	0
Algorithm: Naïve Bayes				
Heart	trestbps	SYS	SYS	0
Diabetes	BloodPressure	DYS	DYS	0
Kidney	bp	DYS	DYS	0
Algorithm: Logistic Regression				
Heart	trestbps	SYS	SYS	0
Diabetes	BloodPressure	PP	DYS	54.1
Kidney	bp	PP	DYS	35.2
Algorithm: Neural Networks				
Heart	trestbps	SYS	SYS	42.8
Diabetes	BloodPressure	PP	DYS	91.9
Kidney	bp	SYS	DYS	98

Table 9. Validation result on different Dataset

The results produced in training and testing phase is validated against dataset of patient suffering from 3 major diseases caused by BP. The datasets used for validation belongs to Diabetes [36], Heart problem [37] and kidney disease [38]. It has 768, 303 and 400 patient records respectively. Only the BP attributes are selected from these datasets. For calculation of accuracy of BP attribute, each dataset is divided into multiple groups of 10 instances each. Statistical features of each group are calculated and validation is performed on each algorithm.

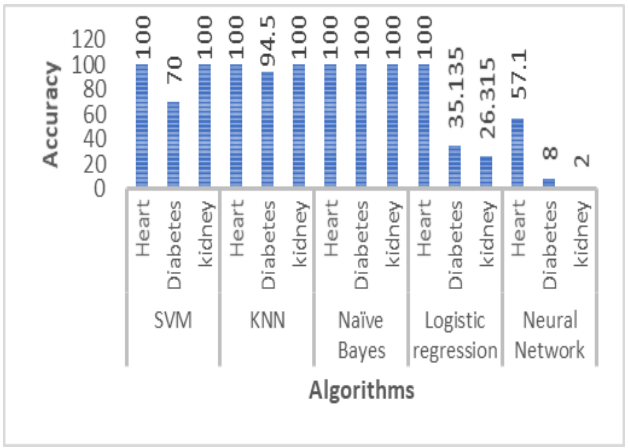

Figure 4. Validation Accuracy results of SMOD on datasets of 3 different Diseases.

Table 9 shows the validation result of each dataset. It shows the attribute label of each dataset as local, predicted and actual. The local attributes labels are those, which is originally labelled and locally used in datasets, the predicted attribute labels are the mapped labels, produced after applying the proposed approach, and actual attributes signify the correct global attribute label. It also shows the percentage of misclassified

data. The table revealed that only the Naïve Bayes algorithm has no misclassified data in all the 3 datasets.

The graph in Figure 4 demonstrates the validation results of SMOD. The result confirms that Naïve Bayes predicts the attribute label with 100% accuracy for all 3 disease datasets while rest of the classification algorithms are not accurate in prediction of atleast one of the diseases. SVM has shown the highest accuracy in training phase but validation results reveal that it is not able to classify all 3 datasets accurately. It clearly shows that, SVM classifier overfits the training data and is not able to pass validation phase. Logistic regression algorithm is able to classify only single dataset as shown. Neural network algorithm fails to predict global attribute with all 3 disease datasets. It shows the highest misclassification results. It is analogous to the less accuracy achieved during training and testing phase.

VII. Conclusion

Health organizations are independently developing and managing their own information systems without adhering to standard guidelines. Data integration and interoperability of patient's physiological records is essential for better analysis of patient health. Schema mapping among the attributes of Polystore databases is essential for data retrieval and its exchange. The research work carried out has proposed a methodology called SMOD for resolving naming conflict in integrating patient's BP data from multiple variety of data sources. The aim of SMOD is to assist in achieving interoperability to make better health decisions. The proposed approach collects BP data from various static and stream data sources and train a multiclass classification model. This model maps local attribute names to its global matching attribute. The mapping enables the accessibility, usability and sharing of Polystore health databases.

Experimental evaluation of SMOD is performed on widely used 5 multi-class classifiers such as SVM, KNN, Gaussian Naive Bayes, Logistic regression and Neural networks. In training phase, the model is compared on the basis of accuracy achieved using 10-fold cross validation technique. In addition to it, comparison of parameters of confusion matrix such as Precision, Recall, F1 score and the average accuracy are used to compare the algorithms.

The results of confusion matrix revealed that SVM and Naïve Bayes algorithms outperforms with accuracy of 100% and Precision 100%, recall 100% and f1-score 100%. The results with 10-fold cross validation also shows the highest accuracy of 99.7% with SVM algorithm. However, Naïve Bayes and KNN both have also performed reasonably well with cross-validation accuracy of 99.1%, Neural network showed less accuracy of 47%, where as its 10-fold cross validation score is approx. 80.5%. The reason of outperforming other algorithms in comparison to neural network is due to the linear nature of datasets, and is and neural network requires large training data to achieve sufficient accuracy as compared to other algorithms.

For generalization of results and to prevent overfitting, the results obtained are validated on BP observation values of three different diseases datasets, caused by BP variation. Such datasets are Heart, Diabetes and Kidney. In validation phase, the results are compared on the basis of percentage of misclassified data and validation accuracy for each dataset. Validation results confirm that only Naïve Bayes algorithm shows 100% accuracy in predicting attribute labels of all the 3 validation datasets. Other algorithms fail to predict labels in atleast one datasets. Naïve Bayes algorithm provides the better result in comparison to other as it distinguishes the two classes based on the highest conditional posterior probabilities. Since, statistical measures are used as training features, there is maximum chances that the unseen test data attribute values are in desired ranges and clearly distinguishes the classes. The possible reason for difference in experimental and validation results is due to the fact that SVM overfit training data while Naïve Bayes algorithm prevent it from fitting too closely.

Thus, Naïve Bayes shows best accuracy in both training and validation phase and is considered as the best and generalized algorithm of schema mapping in the proposed approach. However, the complexity of classification algorithm would be increased when the source dataset linearity is compromised or is more complex, thereby changing the model fit. In general, the existence of noise increases the training time and decreases the accuracy of the classification algorithm but the proposed approach uses statistical computation on dataset, so it would help to normalize the data and thereby reduces the complexity.

The proposed method of semantic mapping has the cardinality matching of 1:1. It implies that single attribute is mapped with only single attribute in another dataset. However, there can be datasets where the other cardinality ratio is also required such as 1: n, m:1 and n: m. To achieve the other cardinalities types, the work can be extended to use multi-label classification algorithms. The proposed approach would also be extended in future for larger datasets distributed remotely and efficient execution of queries over static and streaming data.

References

[1] M. Eichelberg, T. Aden, J. Riesmeier, A. Dogac, and G.B. Laleci. "A survey and analysis of Electronic Healthcare Record standards", *ACM Computing Survey*, 37(4), pp. 277-315, 2005.

[2] A. Salih, M. Salih and A. Abraham. "A Review of Ambient Intelligence Assisted Healthcare Monitoring", International Journal of Computer Information Systems and Industrial Management Applications (IJCISIM), Vol. 5, pp. 741-750, 2014.

[3] Facts About Hypertension | *cdc.gov*. (2020, September 8). Centers for Disease Control and Prevention. https://www.cdc.gov/bloodpressure/facts.htm.

[4] E. Rahm and P. Bernstein. "A survey of approaches to automatic schema matching", *The VLDB Journal*, 10(4), pp. 334–350, 2001.

[5] B. Gu, Z. Li, X. Zhang, A. Liu, G. Liu, K. Zhang, L. Zhao and X. Zhou. "The Interaction between schema matching and record matching in Data integration", *IEEE Transactions on Knowledge and Data Engineering*, 29(1), pp. 186-199, 2017.

[6] M. Birgersson, G. Hansson and U. Franke, "Data Integration using Machine Learning: Automation of Data Mapping using Machine Learning Techniques", In *IEEE Int Enterprise Distributed Object Computing Workshop (EDOCW)*, Austria, pp. 1-10, 2016.

[7] W.S. Li and C. Clifton. “Semantic Integration in Heterogeneous Databases Using Neural Networks”, In *Proc VLDB Conference, Santiago, Chile,* 1994.

[8] S. Sonsilphong, N. Arch-int, S. Arch-int and P. Cherdpan. “A Sematic Interoperability Approach to Health-care data: Resolving data-level conflicts,” *Expert Systems: The Journal of Knowledge Engineering,* 33(6), pp.531-547, 2016.

[9] M. Krol, D. L. Reich and J. Dupont. “Multi-Platforms Medical Computer Systems Integration”, *J Med Syst,* 29(3), pp. 259–270, 2005.

[10] A. Roehrs C. A. da Costa, R. da Rosa Righi, S. J. Rigo and M. H. Wichman. "Toward a Model for Personal Health Record Interoperability", *IEEE Journal of Biomedical and Health Informatics*, 23(2), pp. 867-873, 2019. doi:10.1109/JBHI.2018.2836138.

[11] Y. Yang, X. Li, N. Qamar, P. Liu, W. Ke, B. Shen, Z. Liu. “Medshare: A novel hybrid cloud for medical resource sharing among autonomous healthcare providers”, *IEEE Access*, 6, pp. 46949-46961, 2018.

[12] X. Yang and Y. Miao. “Distributed Agent Based Interoperable Virtual EMR System for Healthcare System Integration”, *J Med Syst*, 35(3), pp. 309–319, 2011.

[13] A. Brugues, S. Bromuri, J. Pegueroles and M. Schumacher. “Providing Interoperability to a pervasive healthcare system through the HL7 CDA Standard”. In *Proceedings of 15th International HL7 Interoperability Conference (IHIC)*, pp. 5-12, 2015.

[14] N. Gupta, and B. Gupta. “Big data Interoperability in e-Health Systems”. In *Proceedings of International Conference on Cloud Computing, Data Science and Engineering (Confluence)*, pp. 217-222, 2019.

[15] C. Kumar, C.V. Rao, A. Govardhan. “A Framework for Interoperable Healthcare Information Systems”, *International Journal of Computer Information Systems and Industrial Management Applications (IJCISIM)*, Vol. 4, pp. 554-561, 2012.

[16] J. Cui, W. Wang, D. Meng, and Z. Liu. “Continuous similarity Join on data streams”. In *Proceedings of IEEE International Conference on Parallel and Distributed Systems*, pp. 552-559, 2014.

[17] P. S. Benny, S. Vasavi and P. Anupriya. “Hadoop Framework for Entity Resolution Within High Velocity Streams”, *Procedia Computer Science*, Vol. 85, pp. 550-557, 2016.

[18] E. Rahm and P. A. Bernstein. “A Survey of Approaches to Automatic Schema Matching”, *The International Journal on Very Large Data Bases (VLDB)*, 10(4), pp.334-350, 2001.

[19] L.S. Wen and C. Clifton. “SEMINT: A Tool for Identifying Attribute correspondence in Heterogenous Databases using Neural Networks”, *Data Knowledge and Engineering*, 33(1), 2000.

[20] P. Mitra, G. Wiederhold, and J. Jannink. “Semiautomatic integration of knowledge sources”. In *Proc. of Fusion’99,* Sunnyvale, USA, 1999.

[21] L. Palopoli, D. Sacca and D. Ursino. “Semi-automatic, semantic discovery of properties from database schemas”. In *Proceedings of International Database Engineering and Applications Symp. (IDEAS), IEEE computation*, pp. 244-253, 1998.

[22] J. Madhavan, P. A. Bernstein, and E. Rahm. “Generic schema matching with Cupid”. In *Proceedings of 27th Int Conf on Very Large Databases*, pp. 49-58, 2001.

[23] R. Fagin, L. M. Haas, M. Hernandez, R. J. Miller, L. Popa, Y. Velegrakis. “Clio: Schema Mapping Creation and Data Exchange”, *Conceptual Modeling: Foundations and Applications*. *Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg, Vol. 5600, pp. 198-236, 2009.

[24] Clinical Knowledge Manager. *openEHR*. URL: https://ckm.openehr.org/ckm/archetypes/1013.1.3574/mindmap [Accessed: Sept., 2019]

[25] HL7FHIR. Retrieved from: (Observation-example-bloodpressure.json - FHIR v4.0.1 (hl7.org)) . [Accessed Sep., 2019].

[26] H. F. Golino, H.F., L. S. de Brito Amaral, S. F. Pimentel Duarte, C. M. Assis Gomes, T. de Jesus Soares, L. A. dos Reis and J. Santos. "Predicting Increased Blood Pressure Using Machine Learning", *Journal of Obesity*, Vol. 2014, Article ID 637635, 12 pages, 2014.

[27] Cardiovascular Disease dataset. (2019, January 20). Kaggle. https://www.kaggle.com/sulianova/cardiovascular-disease-dataset.

[28] P. Muntner, D. Shimbo, R. M. Carey, J. B. Charleston, T. Gaillard, S. Misra, M. G. Myers, G. Ogedegbe, J. E. Schwartz, R. R. Townsend, E. M. Urbina, A. J. Viera, W. B. White, and J. T. Wright. “Measurement of Blood Pressure in Humans: A Scientific Statement from the American Heart Association”, *Hypertension*, 73(5), pp. 35-66, 2019.

[29] M. Nour, and K. Polat. “Automatic Classification of Hypertension Types Based on Personal Features by Machine Learning Algorithms”, *Mathematical Problems in Engineering*, pp. 1–13, 2020.

[30] L. P. Chen, G. Y. Yi, Q. Zhang, W. He. “Multiclass analysis and prediction with network structured covariates”, Journal of Statistical Distributions and Applications 6(1), 2019.

[31] K. Shah, S. Kopru, and J. D. Ruvini. “Neural Network based Extreme Classification and Similarity Models for Product Matching,”. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 3. Association for Computational Linguistics, pp. 8-15, 2018.

[32] A. Doan, P. Domingos and A. Halevy. “Learning to Match the Schemas of Data Sources: A Multistrategy Approach”, *Machine Learning*, Vol. 50**,** pp. 279–301, 2003.

[33] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, D. Brown. “Text Classification Algorithms: A Survey”, *Information*, *10*(4), pp.150, 2019.

[34] P. N. Van Buren, and R.Toto. “Hypertension in diabetic nephropathy: epidemiology, mechanisms, and management”, *Advances in chronic kidney disease*, 18(1), pp. 28–41, 2001.

[35] NCD Risk Factor Collaboration (NCD-RisC). “Contributions of mean and shape of blood pressure distribution to worldwide trends and variations in raised blood pressure: a pooled analysis of 1018

population-based measurement studies with 88.6 million participants", *International Journal of Epidemiology*, 47(3), pp. 872–883i, 2018.

[36] Kaggle. Available from: (https://www.kaggle.com/uciml/pima-indians-diabetes-database). [Accessed Sep., 2019].

[37] UCI. Available from: (https://archive.ics.uci.edu/ml/datasets/Heart+Disease). [Accessed Sept, 2019].

[38] UCI. Available from: https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease [Accessed Oct, 2019].

Author Biographies

Nidhi Gupta is a PhD. Scholar in Department of Computer Science and Engineering at Jaypee Institute of Information Technology, Noida-62, India. She is working as an Assistant Professor in Inderprastha Engineering College, Ghaziabad, affiliated to Dr. A.P.J Abdul Kalam University, U.P, India. She received gold medal in degree of Master of Technology, in 2009 from Banasthali University, Rajasthan, India. Her area of interest are Machine Learning, Data Integration and Interoperability and Big Data.

Dr. Bharat Gupta has received his PhD. Degree from University of Westminster, London, United Kingdom 2005. He is an Assistant Professor in Department of Computer Science Engineering and Information Technology at Jaypee Institute of Information Technology, Noida-62, India. His area of interest are Big data, Machine Learning, Deep learning and Cloud computing.