Privacy Protected Medical Data Classification in precision medicine using an Ontology-based Support Vector Machine in the Diabetes Management System

C. Mallika, S. Selvamuthukumaran

Abstract: Diabetes is a serious health issue across the globe. As stated by the International Diabetes Foundation, currently 425 million people live with diabetes globally, and another 300 million people are expecting to be at higher risk of diabetes in the year 2030.Hence, there is an urgent clinical need for an early prognosis, diagnosis, and management of diabetes and its complications. In this context, numerous intelligent machine learning and data mining approaches including Support vector machine (SVM) have been exploited for diabetes management.SVM is a prevailing supervised and discriminative data classifier which assists the healthcare and biomedical professionals to ascertain unknown data patterns by training a large volume of real-time data. Since this database is the private asset, the sensitive information should be safeguarded without compromising the utility.

Even though SVM is a fast and accurate machine learning technique, it does not have the capacity to represent semantically the classification and reasoning rules which can enable more precise classification. Therefore, the objective of this research is three-fold. Firstly, in order to protect sensitive clinical data, we introduce the Kronecker product and Crow Search Algorithm (CSA) based coefficient generation technique. Secondly, we design diabetes ontology to define domain concepts and relationships and allow medical datamining. Finally, we implement the Ontology-based Support Vector Machine (Ont-SVM) classifier by assimilating privacy protection, ontology, and SVM-based classifier for the diagnosis of diabetes.The experimental results on a real-world dataset, the Pima Indian database at the UCI repository, illustrate that the proposed Ont-SVM outperforms other existing approaches in terms of privacy, accuracy, specificity, and sensitivity.

Keywords: Bigdata; diabetes; Kronecker product; ontology; precision medicine; support vector machine.

I. INTRODUCTION

Healthcare is a thriving domain in many countries that has developed as a predominant sector associated with the increase in employment and revenue as well as physiological burdens [1], [2]. From the analysis of the healthcare cost in India, it is expected that the public medical cost will get doubled in the next five years. Indian healthcare cost is projected to rise from ₹267000 crores in the year 2018 to ₹486000 crores by 2022 [3]. The amount of expenditure on healthcare is projected to increase from 2018 to 2022 to be 45%.Inefficient and non-value adding activities such as incorrect usage of antibiotics, readmissions, and fraud generate a lot of care cost.In India, 5.2 million people die every year owing to medical mistakesand the effects of adversarial processes [4]. An appropriate clinical decision support system (CDSS) can alleviate these issues and enable the transition towards a value-based medical system [5].

Currently, most of the medical organizations are espousing information technology for their organizational venture to provide a better care plan in both technical and commercial aspects [6]. Consequently, a massive amount of data (i.e., terabytes or petabytes) is gathered through this system regularly. The power and effectiveness of big data in healthcare systems, including disease control and health management are derived from its potential of commensurate techniques to intelligently retrieve information and create models from data [7]. Intelligent data mining approaches deliver effective mechanisms to retrieve information from this intricate and vast data and transform all available information into valuable knowledge in order to support a decision-making process.

Diabetes mellitus has been a severe health issue across the globe for several years. Numerous nationwide and worldwide epidemiological research works have found the rising number of diabetic patients in developing and developed countries. Diabetes mellitus is a non-contagious disease pigeonholed by hyperglycemia and related to defects in the metabolism of protein, fat, and carbohydrate. Furthermore, it affects almost every organ such as skin, eyes, heart, kidney, foot, etc. The diabetes management system needs continuous medical care to reduce the risk of enduring complications and to thwart deadly complications.

Sensitive clinical data are supposed to be shared and managed through cloud computing platform where the physician can retrieve information at any time and from anywhere. On the other hand, users of medical sector want to be sure that their personal data are not abused. They want much more control over their sensitive information and they want to be familiar with how their data is exploited, revealed, and safeguarded. They are also concerned about the probable socio economic cost of such data abuse [8], [9]. Patients are reluctant to share their data for other than the treatment purposes and they want to be adequately informed about data sharing. The right to choose

Revised Manuscript Received on November 27, 2019.
* Correspondence Author
C. Mallika, E.G.S.Pillay Engineering College, Email: cmallikachinna@gmail.com
Dr. S. Selvamuthukumaran, A.V.C College of Engineering

what kind of data can be shared under what circumstances establish their privacy rights and need to be employed for privacy and data protection.

A medical diagnosis is a classification process. Data classification is broadly defined as the process of labelling hidden data patterns through a classifier of trained datasets [10]. Owing to its strong ability to differentiate the 'inherent attributes' of various types of data samples SVM is commonly used for the diagnosis of diabetes [11].SVM-based classification contains two stages: (i) training and (ii) testing. During the training process, the classifier identifies the classification measures by training data samples. In the testing phase, the server performs classification based on those measures and tags them to the corresponding class.

This work aims to achieve the classification of diabetes mellitus such that the confidentiality of the sensitive information is protected. The proposed classification technique includes protecting sensitive information by means of Kronecker product-based CSA, which optimizes secrecy-usage (SU) coefficient to protect sensitive information. Afterward, domain ontology is designed to diagnose various types of diabetes by finding clinical similarities between the ontological rules with the trained database. Furthermore, for the classification of diabetes mellitus, Ont-SVM is developed by integrating privacy protection technique, ontology, and SVM-based classification. The performance of Ont-SVM is evaluated in terms of privacy, accuracy, specificity, and sensitivity. The contributions of this research work are three-fold:

- We propose the Kronecker product-based Crow Search Algorithm to protect user privacy.
- We develop domain ontology for diabetes management.
- We devise an ontology-based SVM classification to diagnose different types of diabetes mellitus. Ont-SVM combines privacy protection technique and ontology rules with SVM.

In Section II, we present an analysis of the domain of interest. Section III provides substantial relevant methods aiming to medical data classification with privacy protection. A detailed description of the proposed work is presented in Section IV. Section V depicts the designing of the domain ontology. We discuss the experimental results in Section VI. We conclude this paper in Section VII.

II. DOMAIN ANALYSIS

This section provides the necessary background knowledge of diabetes mellitus. Diabetes mellitus is a chronic and non-contagious disease with complete and/or relative insulin deficit. Diabetes progression is strongly linked with the main irregularities in the metabolism of fat, carbohydrate, glucose, and protein. Additionally, it is accompanied by a notice able tendency to develop neurologic impairment, renal failure, leg amputation, and premature cardiovascular diseases. The diabetes management system needs continuous medical care to avert time delay and expensive treatment. Diabetes can be categorized into four clinical types as given below:

A.Type 1 Diabetes (T1D)

T1D or juvenile diabetes affects children and young individuals (i.e., 18 years old or under). It is mainly caused by the deficiency of insulin due to the β-cells (i.e. insulin-secreting cells) in the pancreas does not generate adequate insulin. As stated in National Diabetes Statics Report [12] T1D affects almost 5-10% of all diagnosed cases globally. People with T1D need external insulin therapy on a regular basis to control blood glucose levels. The progression of T1D is associated with heredity and environmental factors.

B.Type 2 Diabetes (T2D)

T2D is non-insulin dependent diabetes and account 90-95% of diabetes cases identified in adult people [12]. T2D is caused by the fat, insulin resistance, and liver cells which are inadequate to consume insulin efficiently. The risk of T2D complications is related to age, overweight, laziness, unhealthy nutritional habits, history of gestational diabetes, the family history of diabetes, and destruction of glucose metabolism.

C.Prediabetes

It is a milder form of diabetes in which patients have high blood glucose. It is also known as impaired glucose tolerance. In addition, patients with prediabetes have a high risk of having T2D. It can be identified by a simple blood test.

D.Gestational Diabetes mellitus (GDM)

GDM is any degree of glucose intolerance identified during pregnancy that typically vanishes once the pregnancy is over. Based on the National Diabetes Statics Report [12], 5-10% of females with GDMremains to have higher levels of glucose in blood and successively identified as diabetes, generallyT2D. Moreover, the kids of females who had GDM during gestations may be at risk of having over weight and diabetes afterward.

III. RELATED WORK

This section discusses some major contribution to privacy protected data classification with their downsides. Rahulamathavan et al. proposed a classification technique with privacy protection of sensitive information by hiding the classifier from the client during the classification process. Although the efficiency of the proposed approach is the maximum, it has significant computational complexity to perform privacy preserved classification [11].Rizwana Shaikh and Sasikumar suggested a classification approach in the cloud computing environment by introducing a set of constraints that could offer privacy protection depends upon the type of data and accessibility. Though this technique is powerful it failed to formulate a mathematical model for data classification, rather than defining [13].

Chandramohan et al. proposed a privacy-protecting system in the cloud computing environment to avoid digital data loss. The proposed approach helps the clients to protect their sensitive data. Nevertheless, the system is appropriate only when the privacy policy is guaranteed by both data providers and the clients [14].

Kaur and Zandu proposed a method with two levels of computing, namely authentication and storage, to analyze the security issues in the cloud computing environment. In order to facilitate decision making for data classification, several constraints are tobe taken into account, and different techniques are mandatory to enhance the efficiency of classification [15].Farooq and Hussa in presented a hybrid CDSS which consists of two parts:(i) ontology-driven clinical risk assessment and recommendation system, and (ii) machine learning-driven prognostic system of cardiovascular diseases. The proposed CDSS can manage medical errors in the risk assessment process. But, the scientific proof of this approach is limited [16].

Xia et al. proposed a security-based multi-keyword ranked searching approach in cloud computing that was encoded by enabling dynamic update activities such as removing and injecting records concurrently. Although the aim of this approach is to deliver security, there are many security challenges, which need reconstruction of the index [17]. Zardari et al. presented a classification approach based on data confidentiality. A classifier with K-nearest neighbor (KNN) algorithm is implemented to achieve classification according to their security demands. In order to deliver necessary security, sensitive information is encrypted by means of the Rivest-Shamir-Adleman algorithm. The higher computational overhead is considered as the main disadvantage of this approach [18].

Kulkarni and Murugan proposed a new metric, namely c-mixture, for preserving the confidentiality of the data without compromising its utility. A genetic algorithm based on C-mixture was designed by integrating multiple privacy constraints with genetic algorithm for protecting the data before publishing it. The proposed algorithm can resolve the cold-start issue that occurs frequently in the common computational platform. However, considering several parameters in the objective function rises the computational overhead [19]. Fouda et al. developed fuzzy-based domain ontology to predict cardiac arrhythmias that involved diseases, symptoms, diagnosis, and therapies, using Web Ontology Language (OWL). The class hierarchy of the fuzzy ontology is derived from biomedical and disease ontologies. On the other hand, OWL may seem to have compatibility issues and is difficult to train [20].From the literature survey, it is observed that the prevailing privacy protection approaches mainly emphasis on small-sized databases that contain a single attribute to avoid the storage and computational complexity issues. In fact, all the data attributes comprise some confidential information to be preserved. Hence, it is required to take all the attributes into account that have confidential information regardless of the size of the database.

IV. SYSTEM MODEL

This section discusses a schematic representation of our proposed system. The framework consists of three modules: privacy-protection, ontology rule generation, and classification. Figure 1 elucidates an integrated overview of our proposed model.Kronecker product-based CSA is implemented to protect the privacy of the data by generating an optimal SUcoefficient. Ontology is designed by considering the terms of diabetes mellitus used in the healthcare sector. Then, privacy protection and ontology are integrated with SVM for classification.

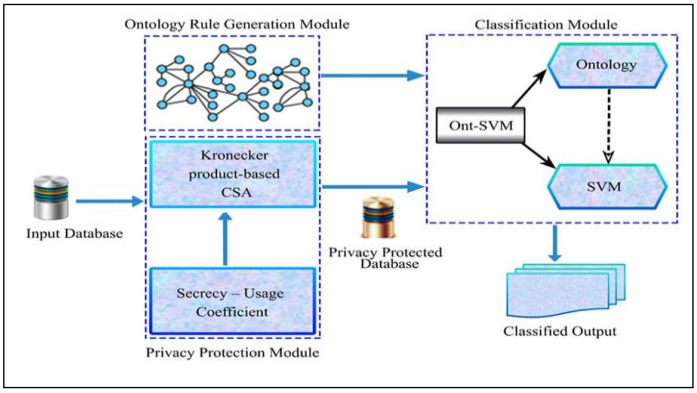


Figure 1. System architecture

A. Kronecker Product

Kronecker product-based CSA is proposed in this work to realize privacy by optimizing SU coefficients. The notion of converting the input data I into privacy preserved output O is expressed mathematically as:

$$O = (C_1 \times S_1) + (S_2 \times C_2) \qquad (1)$$

where C_1 is a correlation matrix of size n×(x∗n), C_2 is a correlation matrix of (k∗m)×m,S_1 is SU-enabled data matrix of size (k∗n)×m, S_2 is SU-enabled data matrix of n×(k∗m),+ is matrix addition, × is matrix multiplication, and O is the privacy preserved output data of size n×m which is exactly the same size of the input data, I and contains n data objects and m attributes. The privacy preserved output is derived by the following steps.

Step 1: Determination of the SU-enabled data matrix S_1

Determine SU-enabled data matrix S_1 by calculating the Kronecker product of the input I and the SU-coefficient matrix, $SU_c$to provide a good trade-off between privacy and utility of sensitive information. The proposed technique absolutely depends on the SU_c since it is the main factor of the data transformation.

$$S_1 = SU_c \otimes I \qquad (2)$$

where⊗ represents the Kronecker product. The coefficient matrix SU_c is of size k×1 which will be optimally selected by means of CSA. For example, if the following numerical values for assigned $SU_c$and I, then the value of S_1 is calculated as follows

$$SU_{c\,[2\times1]} = \begin{bmatrix}0.5000\\0.5000\end{bmatrix} I_{[2\times3]} = \begin{bmatrix}2.8757 & 3.0913 & 2.2865\\9.6721 & 8.9642 & 3.2759\end{bmatrix}$$

$$S_{1\,[4\times3]} = \begin{bmatrix}1.4379 & 1.5457 & 1.1433\\4.8361 & 4.4821 & 1.6379\\1.4379 & 1.5457 & 1.1433\\4.8361 & 4.4821 & 1.6379\end{bmatrix}$$

Step 2: Determination of the data matrix S_2

$S_2$of size n×(k∗m) is determined by finding the Kronecker product of the transposed matrix of $SU_c$and the input data matrix.

$$S_2 = SU_c^T \otimes I \qquad (3)$$

whereSU_c^T is the transposed matrix of the $SU_c$of size 1×k,

$$SU_{c\,[1\times 2]}^T = [0.5000 \quad 0.5000]$$

$$S_{2\,[2\times 6]} = \begin{bmatrix} 1.4379 & 1.5457 & 1.1433 & 1.4379 & 1.5457 & 1.1433 \\ 4.8361 & 4.4821 & 1.6379 & 4.8361 & 4.4821 & 1.6379 \end{bmatrix}$$

Step 3: Determination of the correlation matrixes C_1 and C_2

The correlation matrix of $C_1$with size n×(k∗n) is calculated by correlating every variable from the data input matrix I withS_1.

$$C_1 = Cor(I, S_1) \qquad (4)$$

$$C_{1[2\times 4]} = \begin{bmatrix} 1.000 & 0.9349 & 1.000 & 0.9349 \\ 0.9349 & 1.000 & 0.9349 & 1.000 \end{bmatrix}$$

Similarly, C_2 is calculated by correlating every data attributes of the transposed matrix$S_2^T$with the input data matrix. This correlation provides the matrix of (k∗m)×msince the S_2^Tconsists ofk∗mattributes and input data matrix comprises of m attributes.

$$C_2 = Cor(P_2^T, I) \qquad (5)$$

$$C_{2[6\times 3]} = \begin{bmatrix} 1.000 & 1.000 & 1.000 \\ 1.000 & 1.000 & 1.000 \\ 1.000 & 1.000 & 1.000 \\ 1.000 & 1.000 & 1.000 \\ 1.000 & 1.000 & 1.000 \\ 1.000 & 1.000 & 1.000 \end{bmatrix}$$

Step 4: Determination of privacy preserved output data O

As a final point, the privacy preserved data O is calculated with the same size of the input data using Equation (1).

$$O_{[2\times 3]} = \begin{bmatrix} 20.1721 & 19.7258 & 13.6029 \\ 34.2722 & 33.7665 & 27.3257 \end{bmatrix}$$

B. Crow Search algorithm

The notion behind CSA is to mimic the smart behaviors of crows when stealing and hiding foods which have several resemblances with an optimization process.CSA is a meta heuristic algorithm first introduced by Askarzadeh and has been utilized in many engineering applications especially in the domain of optimization problem [21]. Crows are the most genius birds on earth. They have the highest brain to body ratio. They are capable of identifying human faces and passing information to their families when a hostile one reaches. Also, they can make and employ tools in advanced ways. They remember their food's hiding places (positions)across seasons [22].

Crows have been notorious to follow other birds, detect their food's hiding place, and take it once the proprietor goes away. When a crow has committed stealing, it will take additional precautions like changing their positions to circumvent being a future victim. Actually, crows utilize their own knowledge of having been a thief to envisage the activities of other crows and can find the ideal technique to safeguard their foods from being stolen [23].The basic assumptions of this algorithm are (i) crows live in the flock; (ii) they remember their caches; (iii) they watch each other to steal food; and (iv)they protect their hiding places from being stolen.

Assume that there is a d-dimensional search space. The flock size is N and the position of crow x at time*t* in the search space is definedas $\delta^{x,t}=(\delta_1^{x,t},\delta_2^{x,t},\ldots..\ \delta_d^{x,t})$ where x=1,2,3….N and t=1,2,3…..t_{max}.Here, t_{max} is the maximum iteration. Each crow has a memory where in the best position has been remembered. At time t, the position of crow x is given by $L^{x,t}$. This is the ideal place that crow x has attained hither to. Moreover, crows change their position and discover better food sources. Consider crow y decides to visit its hiding place, $L^{y,t}$. At iteration t, crow x wants to watch crow y to reach $L^{y,t}$. Then two cases can arise.

Crow y is not aware of being observed by crow x.Consequently, crow x will approach to $L^{y,t}$. Then, the new position of crow x is calculated as below:

$$\delta^{x,t+1} = \delta^{x,t} + rand_x \cdot F^{x,t} \cdot (\square^{y,t} - \delta^{x,t}) \qquad (6)$$

Where *rand* is a random number between [0,1] and $F^{x,t}$ represents the flight length of crow x at iteration t.

Crow yis aware of being observed by crow x. Therefore, to safeguard its food from being stolen, crow y will cheatcrow x by moving to a new position. Now, cases 1 and 2 becomes

$$\delta^{x,t+1} = \begin{cases} \delta^{x,t} + rand_x \cdot F^{x,t} \cdot (\square^{y,t} - \delta^{x,t}) & rand_y \geq PA^{y,t} \\ a\ random\ position & otherwise \end{cases} \qquad (7)$$

Where $rand_y$ is a random number and $PA^{y,t}$represents the probability of awareness of the crow y. usually, Meta heuristic algorithms provide a decent trade-off between intensification and diversification [24]. The intensification and diversification factors are primarily regulated by the value of *PA*. By decreasing *PA*, this algorithm tends to performa local search. At the same time, for a small *PA*,the intensification is increased. Conversely, by increasing *PA*, the probability of searching the area of current good solution reduces and this algorithm intended to discover the search space on a global scale. Accordingly, the utilization of large values of *PA* increases diversification.

D.Domain Ontology and rule reasoning

Ontology is an important model used to build clinical guidelines for representing and formalizing medical knowledge in the healthcare system. The main objective of medical ontology is generating a common vocabulary among medical professionals to exchange and reuse knowledge. It explicitly defines the entities or concepts, relationships between the concepts, constraints and their attributes [25].For making a decision about diabetes, ontology is built and then integrated with SVM classifier. The quality of CDSS mainly depends on the completeness and correctness of knowledge that will deduce clinical decisions. Proficient decision support system consists of three phases viz. (i) knowledge acquisition, (ii) semantic modeling, and (iii) knowledge representation

E.Knowledge Acquisition

In order to build ontology, the domain and scope should be described first. The diabetes ontology contains entities for diabetes management system including types of diabetes, symptoms, diagnosis, and treatment. Knowledge acquisition is performed in the following two steps.

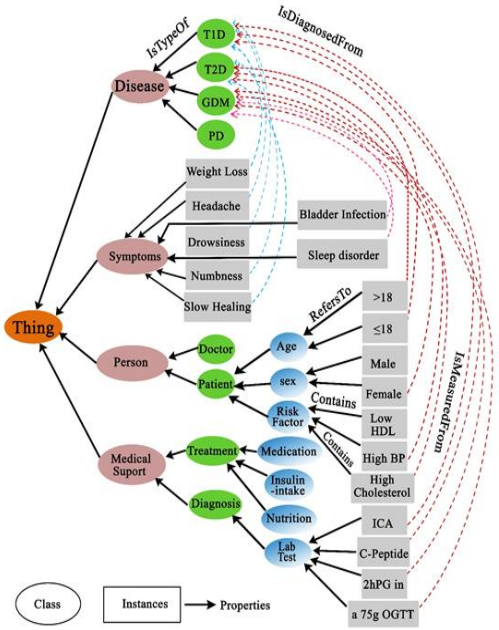
(1) Resource identification: The resources for clinical practice guidelines (CPG)are recognized and used as our references for knowledge procurement. CPG are developed and used by many healthcare organization and regulatory authorities. Though these resources comprise of evidence-based guidelines they dearth a perfect format and hence, are difficult to understand by non-expert end users. After examining these resources US National Guideline Clearinghouse (NGC) [26] and American Diabetes Association (ADA) [27] are adopted in this work.

(2) Conceptualization: In this step, the entities and relationship within the diabetes domain are identified and defined terms used to denote these concepts and their attributes. In the diabetes management system, characteristics of different types of diabetes and symptoms have been identified. To diagnose diabetes, it is essential to acquire patient profile information like age, sex, weight, blood pressure, cholesterol level, etc. and clinical examinations.

The diabetes ontology developed in this work consists of the collection of patients, symptoms, type of diabetes, medical support including diagnosis and treatments, thus making a healing plan based on patient's requirements to construct an efficient CDSS. This framework consists of four classes such as *disease* (holds the characteristics of four types of diabetes), *symptoms* (holds the signs that appear on the diabetic patient), *person* (holds the patient health records and details of medical professionals), and *medical support* (holds diagnosis and treatment recommendation details) as given in Figure 2. These 4 classes have several subclasses and many instances. There area number of attributes in the diabetes management system such as age, sex, weight, blood pressure, body mass index, cholesterol, fasting blood sugar, etc.

The class *disease* contains four subclasses (T1D, T2D, GDM, and PD), and the class *person* has 2 subclasses (patient and doctor), and *medical support* has 2 subclasses (diagnosis and treatment). The subclass *patient* consists of three more subclasses (sex, age and risk factors), where the *age* has 2 instances ((≤18 and >18), *sex* has 2 instances (female and male) and *risk factors* contains 3 instances (low HDL, high BP, and high cholesterol). Similarly, other classes have subclasses and instances as shown in the Figure 2.

Figure 2Class hierarchies for diabetes ontology



F.Semantic Modelling

Semantic modelling represents the knowledge acquired from the previous phase by means of suitable models. In this work, we use production rules to represent diabetes diagnosis system. The rules check the patient's complaints and signs witnessed by the medical professional, risk factors and lab results. Using CPG we frame specialized rules for each type of diabetes. Moreover, it is imperative to devise the rules using a decision tree to identify all type of diabetes and to validate our system properly. For instance, the patient's age is the most significant factor, as most people identified with T1D are less than or equal to 18 years old. Hence, the exclusive characteristic of T1D is that patients are within this age. The general representation of ontology used for the identification of diabetes is given below.

$$\text{ont} := \{E, P\} \qquad (8)$$

Where E denotes a set of entities and P represents the properties of the entity. The properties identified in this work are Is Type Of, Refers To, Contains, Is Diagnosed From, and Is Measured From as given in Table 1.

Table 1. Properties of diabetes ontology classes

According to identified properties, the relationship between the concepts is derived, which relates the classes (x), subclasses (y),and instances (z). The ontological rules can be expressed as $\text{ont}^R = \text{ont}_1^R, \text{ont}_2^R, \text{ont}_3^R \ldots\ldots \text{ont}_n^R$, where *n*is the number of rules formed. Then, the relationship is derived as follows:

$$\text{ont}_1^R = \{x \perp y \perp z) \qquad (9)$$

$$\text{ont}_2^R = \{x \perp z) \qquad (10)$$

$$ont_3^R = \{y \perp z) \qquad (11)$$

$$ont_4^R = \{z_1 \perp z_2) \qquad (12)$$

where⊥denotes the relationship in the ontology. Some examples of these rules are given in Table 2.

Table 2. Examples of ontology rule generation

Rule	Example	Description
1	Disease ⊥ T1D ⊥ ≤18	T1D disease mostly affects children and young adults (≤18 years old)
2	T2D ⊥ Slow healing	Slow healing is a symptom of T2D.
3	GDM⊥ Female	Gestational diabetes affects female during pregnancies

Since every property $k_i \in ont_i^R$, the required property is selected based on the measure *(m)* that assesses the similarity of i[th] property k_i with the rule ont_i^R. The measure *m* is estimated according to the number of features, T shown in Equation (13)

$$m(k_i) = \frac{z^i}{T} \qquad (13)$$

such that $z^i = \sum_{i=1}^{l} z^i$, where $z^i = 1$, if i[th] rule matches the i[th] attribute k_i otherwise it is zero.

$$z^i = \begin{cases} 1 & if\ r_i\ belongs\ to\ k_i \\ 0 & otherwise \end{cases} \qquad (14)$$

When the ideal features are selected according to the score gained in Equation (13), the size of the output data matrix O gets abridged to n × r from n × m, which is the input dataset given to SVM for the classification.

G.Ontology-based SVM

The key idea of SVM is to define a hyper plane that distinguishes training examples in n-dimensional space with the highest margin or interclass distance (h). A cost factor $(\gamma_{+ve}/\gamma_{-ve})$ is introduced as a result of which training errors on positive examples compensate errors on negative examples. Hence, the optimization problem can be expressed as

$$minimize\ \frac{1}{2}\|\eta\|^2 + \gamma_{+ve}\sum_{i:y_i=1} \varkappa_i + \gamma_{-ve}\sum_{j:y_j=-1} \varkappa_j$$

$$subject\ to\ l_k(\eta x_k + h) \geq 1 - \varkappa_k, \varkappa_k \geq 0$$

where η is normal to the hyperplane, $\|\eta\|$ is the Euclidean norm of η, γ represents a factor used top rovide good trade-off between the interclass distance and the training error, $\varkappa_i$ is a slack variable to tolerate classification errors, the class label is denoted by l_i, the perpendicular distance from the hyperplane to the origin is derived from |d|/$\|\eta\|$. In order to manage nonlinearly divisible data,kernel functions and a Lagrange multiplier (β) are introduced. Now the optimization problem becomes

$$maximize\ \eta(\beta) = \sum_{i=1}^{l}\beta_i - \frac{1}{2}\sum_{i=1,j=1}^{l}\beta_i l_i \beta_j l_j K(x_i x_j)$$

$$\gamma \geq \beta_i \geq 0 \qquad \forall_i\ , \quad \sum_{i=1}^{l}\beta_i l_i = 0$$

The training examples with a nonzero β are known as supportvectors(SV)which are used to define the hyperplane. On the other hand, SVMs are black box models. That is, they are not able to provide logical justification for their decisions. For diabetes management system, it is essential to provide the clarification of a decision for the acceptance of black box models by end users. We implement a domain ontology module, which turns the "black box" model of an SVM into more intelligible.

V. RESULT AND DISCUSSION

The performance of the Ont-SVM based classification technique is evaluated by comparing the experimental results with that of four similar techniques, namely (i) Bat Algorithm [28] (ii) PUBAT [29] (iii) CSA [21], and (iv) PP-SVM [11].

A.Experimental setup

For experimentation, we use Intel Core i5 processor with Windows 10 OS. The abovementioned approaches are realized by means of cloud sim tool.The results are obtained by changing the size of the dataset, i.e., N=10 and 20. The dataset for the experimentation is obtained from the National Institute of Diabetes and Digestive and Kidney Diseases. This dataset consists of 768 clinical records of Pima Indian heritage, a population living near Phoenix, Arizona, USA [30]. There are 8 attributes associated with this database (i.e., age, plasma glucose concentration, Body mass index, diastolic blood pressure, Triceps skin fold thickness, 2-hour serum insulin, number of times pregnant, and Diabetes nutrition function).Even though the database is considered as there are no missing values, there are some copiously added zeros as missing values. 192 patients had skin fold thickness readings of zero, 28 had a diastolic blood pressure of zero, 11 more had a body mass index of zero, 140 others had serum insulin levels of zero, and 5 patients had a plasma glucose concentration of zero value. After eliminating the records with missing data values, 460 records were analysed for performance evaluation.

B.Performance measures

The classification approaches are evaluated in terms of privacy, accuracy, sensitivity, and specificity. These performance metrics are required to be higher to increase the efficiency of the algorithms. Privacy is calculated from the cosine similarity metric. This measure is used to calculate the similarity between the input and output vector based on the cosine angle between them. Moreover, this metric is to be in minimum which signifies that the data confidentiality is protected. Therefore, the similarity measure is subtracted from unity.

$$Privacy = (1 - \cos(I, O) \qquad (15)$$

The utility is estimated in terms of the accuracy as follows

$$\text{Accuracy} = \frac{TN + TP}{TN + TP + FN + FP} \quad (16)$$

Specificity and sensitivity define how well the classification approach differentiates positive and negative classes. Specificity represents the false alarm rate and sensitivity is the detection of disease rate.

$$\text{Sensitivity (TP rate)} = \frac{TP}{TP + FN} \quad (17)$$

$$\text{Specificity (FP rate)} = \frac{FP}{TN + FP} \quad (18)$$

C.Performance analysis

The percentage of the dataset is varied from 50% to 100% for two different flock sizes (i.e., N=10 and 20). In the analysis with N = 10, the BAT, PUBAT, CSA and PP-SVM have the privacy of 3.2%, 49.9%, 50.0%, and 55.3% for 50% data, whereas Ont-SVM has a value of 73.3%. The privacy decreases in all the comparative techniques when the data rises to 100% as shown in Figure 3(a). The similar trend is observed for N=20 as shown in Figure 3(b). Meanwhile, Ont-SVM generates higher average privacy values.

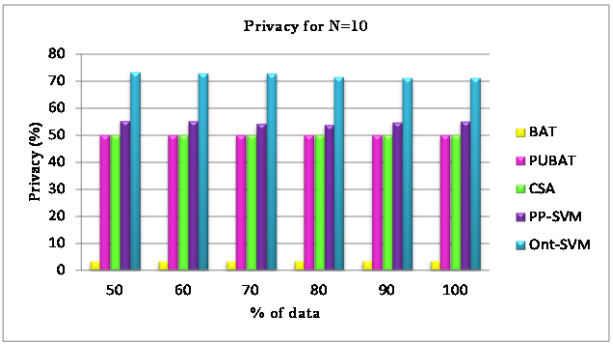


Figure 3 (a) Privacy for N=10

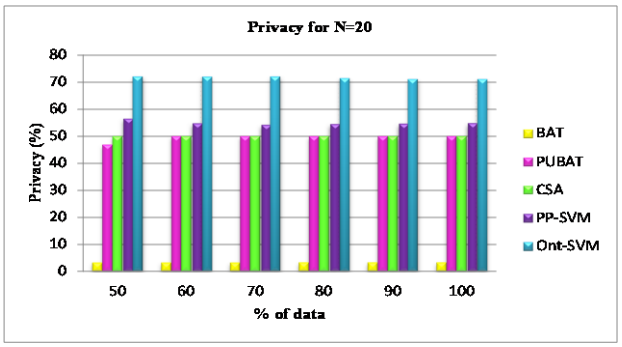


Figure 3 (b) Privacy for N=20

The accuracy of all the approaches increases as the percentage of data is increased. The accuracy analysis for the given dataset for different flock size is shown in Figures 4(a) and 4(b). For 50% data, the accuracy of BAT, PUBAT, CSA, PP-SVM and Ont-SVM is 46.3%, 48.5%, 62.5%, 78.7% and 80.9%. For 100% data the value of accuracy is 46.4%, 49.7%, 55.1%, 78.5% and 81.2%. From figures 4(a) and 4(b), it is observed that the Ont-SVM outperforms other existing approaches in terms of accuracy.

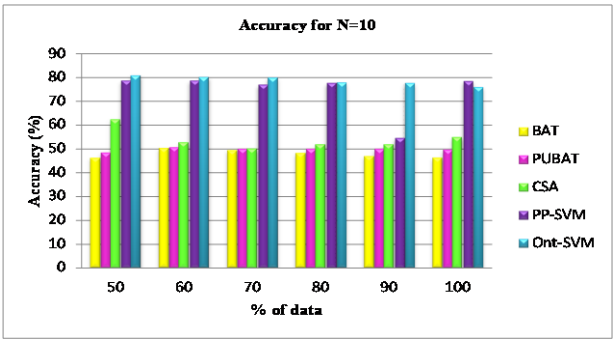


Figure 4 (a) Accuracy for N=10

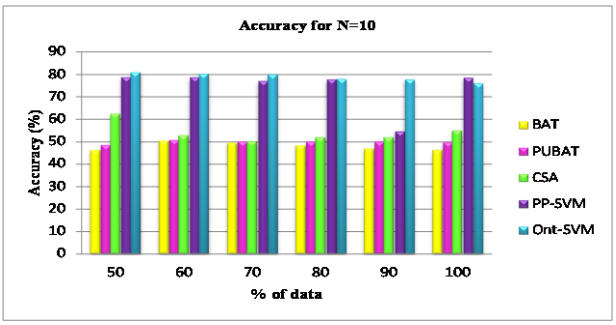


Figure 4 (b) Accuracy for N=20

The performance analysis of all the approaches in terms of sensitivity is given in Figures 5(a) and 5(b). The sensitivity increases as the percentage of data is increased from 50 to 100. Figure 5(a) presents the sensitivity plot for N=10, where the maximum sensitivity obtained from Ont-SVM is 92.4%. The sensitivity obtained in BAT, PUBAT, CSA, and PP-SVM is 83.7%, 86.1%, 87.4%, and 91.9%, when the percentage of data is 100. The sensitivity seems to be increasing in Ont-SVM approach from 91.1% to 93.2%, when the data is varied from 50 to 100%. The similar trend is observed of flock size 20.

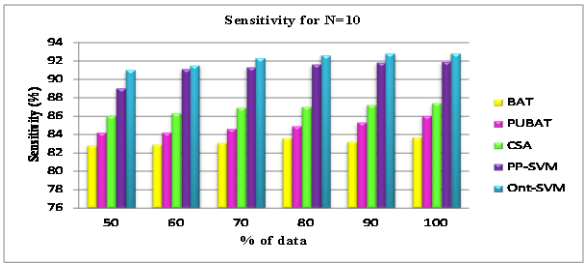


Figure 5 (a) Sensitivity for N=10

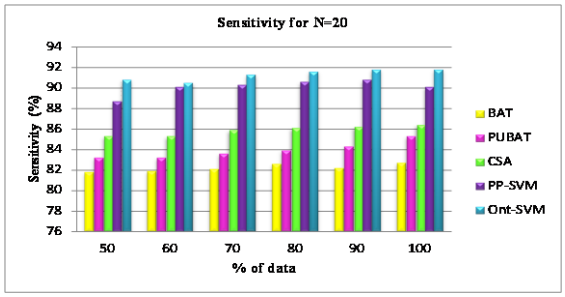


Figure 5 (b) Sensitivity for N=20

The results of the analysis in terms of specificity for different techniques are illustrated in Figure 6(a) and 6(b). Figure 6(a) shows the specificity for N=10. When the input data is 50%, the specificity measured by BAT, PUBAT, CSA, and PP-SVM is 66.23%,

67.59%, 69.35%, and 71.9% respectively while Ont-SVM gives 74.23% specificity. The specificity increases with data percentage. When the percentage of data is 100, the specificity is observed to be in maximum in all the five approaches as shown in figures.

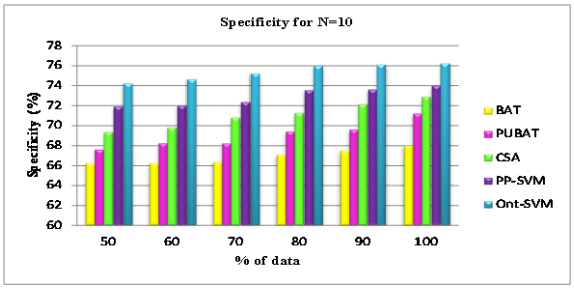


Figure 6 (a) Specificity for N=10

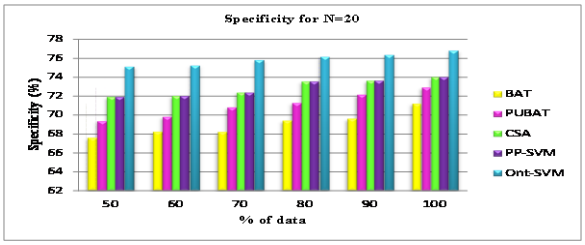


Figure 6 (b) Specificity for N=20

VI. CONCLUSION

Cloud computing provides potential solutions for addressing big data storage challenges. Nevertheless, the cloud servers disclose the patient's medical record without their consent. Hence, it is necessary to develop efficient techniques to protect sensitive clinical information. The research goal of this article is to perform privacy protected data classification for diagnosing diabetes. For this purpose, we introduce the Kronecker product with CSA to protect the confidential medical records. Then, we design domain ontology to define domain concepts and relationships and allow medical data mining. Finally, we implement the Ontology-based Support Vector Machine (Ont-SVM) classifier by assimilating privacy protection, ontology, and SVM-based classification approaches for diagnosis of diabetes. The experimental results on a real-world dataset, Pima Indian diabetic database at the UCI repository, illustrate that the proposed Ont-SVM achieves better performance in terms of privacy, accuracy, specificity, and sensitivity as related to the other existing approaches in the literature. The dataset analysed in this work has some missing values. In our future research, we plan to develop more intelligent data classifier to consider these missing values also.

REFERENCES

1. Yang,J.-J.;Li,J.; Mulder, J.; Wang,Y.; Chen,S.; Wu,H.; Wang,Q.; Pan,H. Emerging information technologies for enhanced healthcare. Comput. Ind. 2015, 69, 3–11.
2. Ali Z, Hossain MS, Muhammad G, Sangaiah AK. An intelligent healthcare system for detection and classification to discriminate vocal fold disorders. Future Gen Comput Syst. 2018;85:19–28.
3. Ministry of Health & Welfare, G. (2017). Malaria : National Vector Borne Disease Control Programme (NVBDCP)
4. Sandip Patel, Ashita Patel, Varsha Patel, Nilay Solanki, Study of Medication Error in Hospitalized Patients in Tertiary Care Hospital, Indian Journal of Pharmacy Practice, Vol 11, Issue 1, Jan-Mar, 2018.
5. Berwick,D.M.; Hackbarth,A.D. Eliminating waste in US healthcare. J.Am. Med. Assoc. 2012,307,1513–1516.
6. Prokosch, H.-U.; Ganslandt, T. Perspectives for medical informatics. Methods Inf. Med. 2009, 48, the abundance of data 38–44.
7. B. Suseela, V. Jeyakrishnan, A multi-objective hybrid aco-pso optimization algorithm for virtual machine placement in cloud computing, Int. J. Res. Eng. Technol.3(4)(2014)474–476.
8. G. Bansal, F. Zahedi, and D. Gefen, "The impact of personal dispositions on information sensitivity, privacy concern and trust in disclosing health information online," Decision Support Systems, vol. 49, 2010, pp. 138-150.
9. R. Au and P. Croll, "Consumer-Centric and Privacy Preserving Identity Management for Distributed E-Health Systems," Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008), 2008, pp. 234-234.
10. Li Z, Zhang L, Zhong R, Fang T, Zhang L, Zhang Z. Classification of Urban Point Clouds: A Robust Supervised Approach With Automatically Generating Training Data. *IEEE J Sel Top Appl Earth Obs Remote Sens*. 2017;10(3):1207-1220.
11. Rahulamathavan Y, Phan RCW, Veluru S, Cumanan K, Rajarajan M. Privacy-Preserving Multi-Class Support Vector Machine for Outsourcing the Data Classification in Cloud. *IEEE Trans Dependable Secure Comput*. 2014;11(5):467-479.
12. Centers for Disease Control and Prevention. (2014). National diabetes statistics report: estimates of diabetes and its burden in the United States, 2014. Atlanta, GA: US Department of Health and Human Services,[Online]. Available at:http://www.cdc.gov/DIABETES/data/statistics/2014 StatisticsReport.html.
13. Rizwana Shaikh, Dr. M. Sasikumar, "Data Classification for Achieving Security in Cloud Computing," Procedia Computer Science, International Conference on Advanced Computing Technologies and Applications (ICACTA), Vol. 45, pp. 493-498, 2015.
14. Chandramohan,D.,Vengattaraman,T.,Dhavachelvan,P.:Asecure data privacy preservation for on-demand cloud service. J. King Saud Univ.-Eng. Sci. 29(2), 144–150 (2017).
15. Kaur K, Zandu V. A Secure Data Classification Model in cloud computing using Machine Learning Approach. Int J Mod ComputSciAppl (IJMCSA). 2016;4:4.
16. Farooq K, Hussain A. A novel ontology and machine learning driven hybrid cardiovascular clinical prognosis as a complex adaptive clinical system. Complex Adaptive Systems Modeling. 2016;1-21.
17. Xia Z, Wang X, Sun X, Wang Q. A Secure and Dynamic Multi-Keyword Ranked Search Scheme over Encrypted Cloud Data. IEEE Trans Parallel Distrib Syst. 2016;27(2):340-352.
18. Zardari MA, Jung LT, Zakaria N. "K-NN classifier for data confidentiality in cloud computing," 2014 International Conference on Computer and Information Sciences (ICCOINS), Kuala Lumpur, pp. 1-6, 2014.
19. Kulkarni, Y.R., Murugan, T.S.: C-mixture and multi-constraints based genetic algorithm for collaborative data publishing. J. King Saud Univ.-Comput. Inf. Sci. (2016)
20. Fouda H, Elmogy M, Aboelfetoh A, Maat AR. "Constructing fuzzy ontology for cardiac arrhythmias," 2015 Tenth International Conference on Computer Engineering & Systems (ICCES), Cairo, pp. 402-409, 2015.
21. Askarzadeh A. A novel metaheuristic method for solving constrained engineering optimization problem: crow search algorithm. ComputStruct 2016;169:1–12.]
22. Prior H, Schwarz A, Güntürkün O. Mirror-induced behavior in the magpie (picapica): evidence of self-recognition. PLoSBiol 2008;6(8):e202.
23. Clayton N, Emery N. Corvide cognition. CurrBiol 2005;15:R80–1.
24. Yang XS. Metaheuristic optimization. Scholarpedia 2011;6. 11472.
25. Tu, S. W., &Musen, M. A. (2001). Modeling data and knowledge in the EON guideline architecture. Studies in health technology and informatics, vol. 84, pp. 280-284.
26. National Guideline Clearinghouse (NGC), Guidelines, www.guideline.gov, [Online]. Available at: http://www.guideline.gov/about/index.aspx
27. American Diabetes Association. (2014). Standards of Medical Care in Diabetes—2014. Diabetes Care, vol. 37, no. Supplement 1, PP. S14-S80. Available at:

http://care.diabetesjournals.org/content/37/Supplement_1/S14.extract
28. Yang, X.S.: A new metaheuristic bat-inspired algorithm. In: Gonzalez, J.R. (ed.) Nature Inspired Cooperative Strategies for Optimization (NISCO 2010), Studies in Computational Intelligence, vol. 284, pp. 65–74. Springer, Berlin (2010)
29. Karlekar, N.P., Gomathi,N.: Kronecker product and bat algorithm based coefficient generation for privacy protection on cloud. Int.J. model. Simul. Sci. Comput. 8(4), 1750021 (2017)
30. UCI repository of bioinformatics Databases, Website: http://www.ics.uci.edu/~mlearn/MLReposit ory.html.

AUTHORS PROFILE

Mrs.C.Mallika is currently working as Associate professor in E.G.S.Pillay Engineering college Scholar and Part time research scholar of Anna university .She has rich teaching experience of 14 Years .she has published 5 papers in International Journals and availed TNSCST FDP grant.

Dr.S.Selvamuthukumaran is currently working A.V.C College of Engineering, Mannampandal, Mayiladuthurai as vice principal and Professor of computer Applications .He posses Ph.d Degree in computer science and has rich teaching experience of 23 Years and research experience .He has Guided 2 research scholars and is presently guiding 6 research scholars .He published 19 papers international journals. He has obtained funds under MODROBS scheme and one seminar and Two FDTP Grant.