

# MapReduce-iterative support vector machine classifier: novel fraud detection systems in healthcare insurance industry

Jenita Mary Arockiam, Angelin Claret Seraphim Pushpanathan

Department of Computer Science, College of Science and Humanities, SRM Institute of Science and Technology, Tamil Nadu, India

---

## Article Info

### Article history:

Received Nov 30, 2021

Revised Jul 14, 2022

Accepted Aug 19, 2022

---

### Keywords:

Big data

Fraud detection

Insurance claims

Iterative support vector machine

MapReduce framework

---

## ABSTRACT

Fraud in healthcare insurance claims is one of the significant research challenges that affect the growth of the healthcare services. The healthcare frauds are happening through subscribers, companies and the providers. The development of a decision support is to automate the claim data from service provider and to offset the patient's challenges. In this paper, a novel hybridized big data and statistical machine learning technique, named MapReduce based iterative support vector machine (MR-ISVM) that provide a set of sophisticated steps for the automatic detection of fraudulent claims in the health insurance databases. The experimental results have proven that the MR-ISVM classifier outperforms better in classification and detection than other support vector machine (SVM) kernel classifiers. From the results, a positive impact seen in declining the computational time on processing the healthcare insurance claims without compromising the classification accuracy is achieved. The proposed MR-ISVM classifier achieves 87.73% accuracy than the linear (75.3%) and radial basis function (79.98%).

*This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.*



---

## Corresponding Author:

Angelin Claret Seraphim Pushpanathan

Department of Computer Science, College of Science and Humanities, SRM Institute of Science and Technology

Kattankulathur, Chennai District, Tamil Nadu 603203, India

Email: angelins@smist.edu.in

---

## 1. INTRODUCTION

The recent advancements made in communication and digital technologies have revolutionized the modern world. It develops a highly connected environment among the communication entities. Different types of networks such as social platforms, e-commerce, blogs, industrial trading, banking and insurance networks are increasing along with the development of communication technologies. A tremendous volume of data is being generated from these networks. A billion transactions are carried out in a fraction of seconds. A vast array of information is easily accessible by the fraudsters (or) attackers via creating anonymous platforms [1]. The growth of anomalous networks, fraudsters have developed several opportunities to manipulate the data without the user's knowledge. Many organizations employ preventive measures to secure their networks and data from internal and external threats with the help of digital technologies. Special considerations are taken on the interactions and the activities performed among the inter-network entities [2]–[5].

A widespread of machine learning (ML) algorithms is incessantly explored in the different fields of real-time applications. In recent years, it has been increasing prominence due to the popularity of big data [6]–[8]. The problems in ML algorithms are known to be the issue of learning from experience by analyzing some tasks and performance measures. It helps the users to unleash the data structure and develop the

predictions model from large datasets. ML proliferates the learning algorithms, rich set of information and dynamic computing environments.

Figure 1 illustrates the role of ML algorithms in big data. The ML component is surrounded by four elements: big data, system, user, and domain. The communication flows between all elements are bi-directional. Large and complex financial data is given as the input to the machine learning components, and then the extensive data computation becomes a part of big data.

Here, the user provides domain analysis and feedback to the learning element, which eases the decision-making process. The domain component provides the context guidelines to the learned models. The System component deals with the infrastructure module that illustrates the usage of computing environments like distributed computing, and edge computing [9].

The growth of assisting connected devices and communication technologies has developed a passage for stealers to manipulate the data, leading to a severe financial loss crisis for the healthcare sectors. Hence, the researchers have explored data security analytics [10]. In the insurance sector, fraud activities distress both customers as well as insurers. It decreases the trust and loyalty between customers and insurers. A diversified process and products in healthcare services are being designed with the use of medical technologies. Healthcare insurance management systems administer the different insurance companies and the healthcare organizations in the marketplace. Generally, it includes two models, namely, the payment model and the claim management models. The real challenge pertains to the claim management process, which allows for more advanced analysis like fraud management [11], [12]. Figure 2 illustrates the scope of the fraud detection approach during the insurance claim management process.

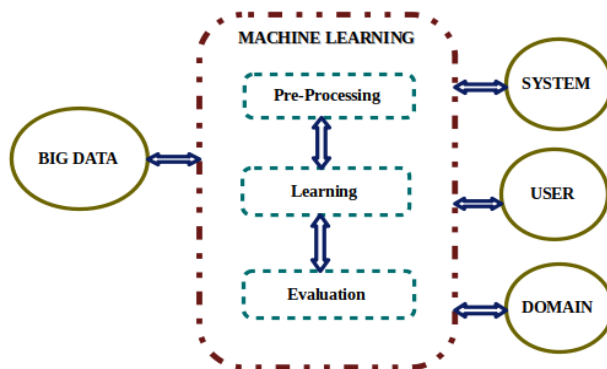


Figure 1. Role of machine learning (ML) in big data

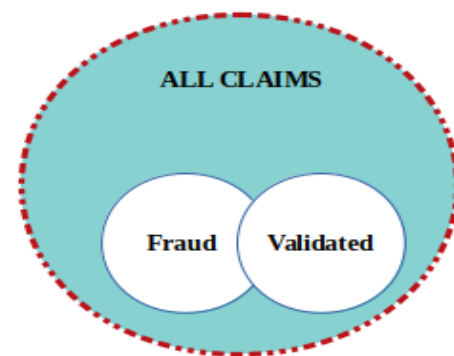


Figure 2. Scope of fraud detection approach

There has been a myriad of studies in detecting fraud claims in the healthcare industry. The review study has analyzed from two aspects, class imbalance and features representation. Class imbalance and feature representation are some of the classical problems in machine learning algorithms [13]. Due to the improper definition of features, an imbalance occurs between the classes, i.e. one class has high data samples whereas the other class has low data samples. Finding the abnormal claims is a challenging task due to the issues mentioned earlier.

The performance of ML techniques in financial frauds has been surveyed by [14]. The general techniques involved in machine learning are descriptive techniques, predictive techniques, artificial intelligence techniques and hybrid techniques. The analysis has stated that the hybrid techniques, genetic algorithm and support vector machine (SVM) outperformed better than other techniques [15]–[17]. The solution to financial frauds is always a never-ending task because of class imbalance. Frauds and abuse are the two factors that incline healthcare costs. Due to class imbalance, it is highly affected by the Brazilian Health Care Market [18]. Service Providers were asked to find out the link between fraud and abuse on the claim authorization process. The execution of cross-validation distribution on treatment methods, machine learning algorithms like SVM, C4.5, random forest and naive Bayes were analyzed. The results have stated that the random forest was not affected by the class imbalance. Other ML algorithms were involved when the class distribution changed.

Insurance frauds have become more complex due to the accumulation of prominent data resolved by big data predictive modelling [19]. Distributed and the ML algorithms tested parallel computing tools to differentiate the fraud records. The fraudulent patterns change over time, and thus, an imbalance between the detected patterns creates trouble for the detection approaches. Concept drift [20] is a domain that

encompasses the dynamic data, i.e. change over time. Labelling of unsupervised data requires concept learning. The authors have dealt with automatic labelling of unsupervised data using the concept drift approach. A permutation test was conducted over each statistical data, and the p-value determined the class of data. Since it follows a one-fixed algorithm, the effects of class imbalance are high in noisy data.

The decision support tools require an intensive SME analysis when it comes to prepayment and post-payment control models. The presence of outliers in medical data lowered the accuracy of the detection framework. Outlier detection techniques [21] were explored to find the misclassified patterns, i.e. false positive rate. It tested on Medicaid data of 650,000 healthcare claims and 369 dentists of one state. An improper cluster formation has increased the FPR, and also, the estimated clusters mean improper class distribution. The comprehensive services provided by healthcare sectors have become more portable by adopting android technologies [22]. The tracking of claiming benefits requires a timely prediction. Because of the complex data granularity, it has lowered the accuracy of the framework. Thus, methods such as semi supervised isomap (SSIsomap) activity clustering, simple local outlier factor (SimLOF) outlier detection, and the Dempster-Shafer theory-based evidence aggregation are studied on the real-world dataset. The behavior profile pattern also alters when the data size increases, which strongly induces the estimated frequent itemset.

The provider-consumer model incurs a considerable expense from the healthcare systems. Thus, the anomaly was detected from the provider and consumer models [23]. Brazilian healthcare records from 2008 to 2015 were collected and evaluated using bipartite graphs and k-nearest neighbors algorithm (k-NN) algorithms. The bipartite charts were employed to find the relationships between those two models. The detected similar patterns used to classify into potential providers and anomaly classes using k-NN. The performance measure cost and effectiveness validate it. Instead of validating the number of hospitals, the available cities and consumer scores was used for effectiveness estimation. Therefore, representing the features are essential in graph-based approaches.

Several researchers have explored Medicaid to discover fraud in medical data beyond the transaction level [24]. The multidimensional data analysis was designed for fraud classifications using sparrow's insights. The discovered fraud patterns are also from unsupervised data. The data was classified into six classes based on the levels of fraudulent data patterns were identified. It was concluded that the inefficiency of training data had lowered the performance of supervised ML techniques. With the above as a base, an ML model was designed to detect frauds done by physicians [25]. When it comes to billing procedures, the frauds may be external (or) internal frauds. Irrespective of the claimer, the physicians were also performing the misuse of billing procedures, which is challenging. Hence, a multinomial Naïve Bayes algorithm was designed to resolve multi-class classification by following 5-cross validation. The classification was done by interchanging the features, like field experts, specialty, and provider types. Relied upon F-score, the fraud levels on procedures done by physicians were reported. It has built an association among different levels of physicians when handling the claim data.

Association rule mining is also employed to recognize fraudulent patterns. It is used to constructing associations/correlations between features. Initially, the transaction data was transformed into a set of clusters, and then some standard association rules [26] were framed. Based on the lift and confidence value, the data samples were classified into fraudulent and non-fraudulent claims. The analysis of claim data was concentrated on the feature extraction phase rather than the classification phase. However, some features are discarded in terms of big data analytics. The invasion of variant actors and commodities [27] in the healthcare insurance claims has imposed different challenges to the ML techniques. Therefore, an interactive framework for unsupervised data analysis was required using pairwise computational models such as analytical hierarchical processing (AHP) and expectation-maximization (EM). CGM Turkey for private insurance companies was validated under area under curve (AUC). It has been stated that the independent analysis of actors and the commodities reduced the time rate for predicting fraud. The fragmented nature of feature representation has brought significant changes towards the facts finding the process of institutions.

The patient rule induction method (PRIM) [28] was designed to extract the anomalies patterns under big data context. It was implemented in Center for Medicare Services (CMS) 2014 dataset, which has improved the feature space. While partitioning the feature space, a depth-analysis on different classes is not done. Since it performs conditional probability on features, the activities of physicians are not traced. Heuristics approaches on defining optimal fraud indicators are not possible due to the higher accumulation of false claims. Fake billing frauds are available more than other frauds, especially in auto/vehicle insurance claims [29]. Comparison models were designed using random forest, naive Bayes and decision tree under confusion matrix measure. It was implemented in a synthetic dataset, which concluded that the random forest has outperformed better than the other two models. Feature modelling has a significant part in designing the classifiers to reduce false positive and true negative rates. Analyzing camouflage behaviors [30] is a troublesome task from the classification approaches because it sustains for a short period. Patient cluster divergence-based healthcare insurance fraudster detection (PCDHIFD) was designed to classify the

fraudulent caused by camouflage behaviors [30]. With the help of patient admission date features, the correlation value between the patients, hospitals and the providers were computed from a graph-based dense peak clustering approach. Then, a divergence cluster value was used to detect the fraud patients. The f-measure has been improved by 15% than other classification models. Interpreting the medical admission-oriented features affects the classifiers in the camouflage behaviors analysis. This research study proposes a novel fraud detection model by hybridizing the strengths of big data and machine learning approaches to solve the insurance claim classification. It reduces the effects of class imbalance over the voluminous data that has multi-classes. The insurance claim data is preprocessed using MapReduce framework that scales up the efficiency of data processing capabilities. The deployment of iterative support vector machine (ISVM) classifier on processed data helps to classify the fraud providers by executing the pointed iterative conditions. The proposed MapReduce based iterative support vector machine (MR-ISVM) classifier achieves the objectives of classification accuracy with the less computational time.

## 2. METHOD

Class imbalance and feature modelling are mutually dependent on supervised based ML techniques. Multi-class learning (MCL) is a challenging domain between ML and big data analytics. The research on MCL has not been suggested more than single-class learning (SCL). MCL is defined as the problem of associating an instance with more than one class, even for binary labels. Conventional methods do not support MCL because it reduces the prediction accuracy of the application framework. MCL requires a systematic approach to handle the medical data effectively and enhances cost-saving and detection efficiency. Feature selection technique (FST) has to take the categorical, continuous, and high-dimensional data to innovate the MCL domain. Let us define the problem in vector form. Each instance in the database is represented as,  $a = \{a_1 \dots a_p\}$  where,  $p$  represents the final instance. The data instance is obtained from the domain,  $D = \{A_1 \dots A_p\}$ . Then, each instance  $a$  is associated with the class labels, is denoted as,  $Label\ l = \{l_1 \dots l_q\}$  where  $q$  represents the final class value. The class labels are obtained from the domain,  $Class_{domain}(C) = \{L_1, L_2 \dots L_Q\}$ . Each class label contains a possible set of class variables  $j$ , which is represented as  $C_j = (1, 2, \dots K_j)$ .

This research aspires on framing a fraud detection model that detects the mishandling of the claiming process using machine learning algorithms. Ideally, the proposed method is designed to discover provider abuse by analyzing the variables used in treatment, disease and claim. The steps of the proposed process are explained in a detailed manner. Figure 3 presents the block diagram of the proposed research. The proposed research comprises five phases, and they are explained in brief:

- Data acquisition: It is the foremost step that portrays the information of datasets.
- Data preprocessing: It is the second step that portrays the organization of the collected datasets.
- Feature selection: The third step describes the selection of features used for constructing the training classifier.
- Classification: It is the fourth step that presents the workflow of the proposed classifier.
- Detection: It is the final step that assists the testing data.

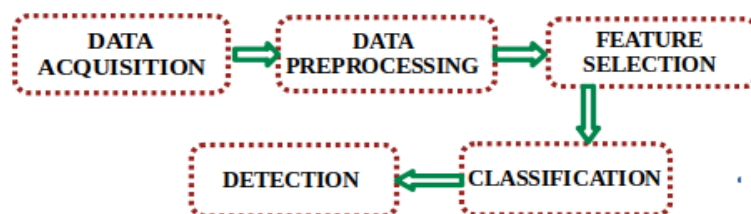


Figure 3. Block diagram of the proposed study

### 2.1. Data acquisition

Dataset is collected from the well-known public repository, known as “Healthcare provider fraud detection analysis” [31]. Provider fraud is one of the biggest scams prevailing in the healthcare industry. Due to the mishandling of disease and the treatment details by the physician, the providers increase the medical costs. The metadata of the dataset is presented in section 3. The collected dataset determines the success rate of the research objectives.

**2.2. Data preprocessing**

This step’s task is to organize the data presented in the datasets efficiently. It is achieved by eliminating the missing values, duplicate data and also developing efficient data partitioning. The examination of missing values and duplicate data is described in the next section. The development of the data partitioning is explored by using a novel MapReduce technique. It is found that multiple claim IDs are generated with various providers, which is differentiated by the diseases. Owing to this, the MapReduce technique is employed over the ‘inpatient and outpatient’ tables. Based on the disorders, a new table is created. As the name suggests, the MapReduce technique consists of viz, mapper, and reducer functions. The mapper function is expressed as,

$$Mapper: (k_1, v_1) \rightarrow [(k_2, v_2)] \tag{1}$$

The reducer function is expressed as,

$$Reducer: (k_2, |v_2|) \rightarrow [(k_3, v_3)] \tag{2}$$

where,  $k_1$ & $k_2$  are the input key and the output key;  $v_1$ & $v_2$  are the input value and the output value;  $(k_3, v_3)$  are the final output key and the value obtained from the reducer function; and  $|v_2|$  is the final data list.

The Figure 4 presents the workflow of the MapReduce technique. It consists of four functions, namely, splitting, mapping, partitioning and reducing. Both the functions execute parallel on the input datasets by creating many subsets under different cluster nodes. The intermediate output values of the mapper function will serve as the input to the reducer functions. Based on the user-defined values of many partitions (p) and the partitioning function, the MapReduce technique executes the steps: i) a unique processor is created for the master and the slave nodes; ii) master nodes are responsible for assigning the task to the nodes in the mapper and reducer jobs; iii) for the user-defined partition values, each partition runs on the mapper node; iv) the output values, i.e. keys and the intermediate values of a mapper job, are preserved in the local files of local storage; v) the keys and the intermediate values on the local files are then assigned to the reducer job; and vi) after completing the reducer job, the reduced output with the final values is stored at the master node.

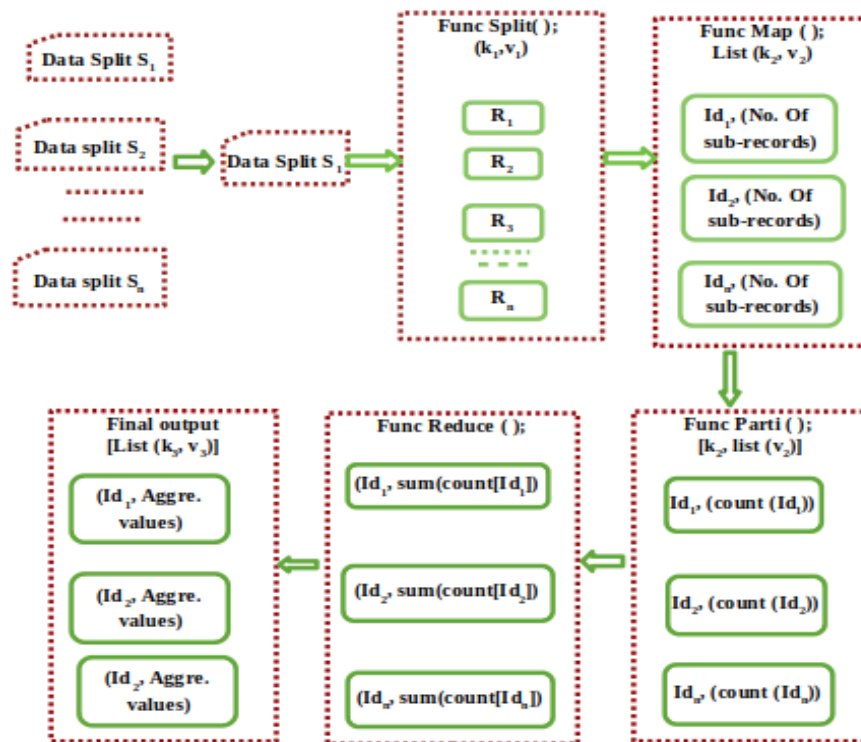


Figure 4. Workflow of the MapReduce technique

### 2.3. Feature extraction and selection

Feature selection is the third step that deals with the extraction of required features to build an efficient training classifier. The data table contains a high set of features, and thus, the importance of each feature is studied to eliminate the irrelevant features. Linear discriminant analysis (LDA) is performed over the data table. The objective of LDA is to explore the linear combination of features that combines two (or) more classes of objects. The most desired features are obtained from reducing the dimensionalities before building the classifier. It is the most suitable model for preserving the multiple classes with reduced dimensions. The claiming procedure depends on the different aspects of the medical reports of the patients. It is found that a beneficiary holds multiple claiming strategies for multiple diseases. The amount is claimed based on the disease code, treatment code and the total amount. Here, three types of variables, viz, claiming variables, disease variables and treatment variables. In this step, we intend to find out the ‘confidence score’ of the bills given by the provider. The estimated confident score will help verify the attributes taken for creating, validating and verifying the statements provided by the provider. The confidence score function is calculated:

$$Z = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_d x_d \quad (3)$$

$$Confidence_{score}^{\beta} = \frac{\beta^T \mu_1 - \beta^T \mu_2}{\beta^T C \beta} \quad // \text{ score function of a class} \quad (4)$$

$$Confidence_{score}^{\beta} = \frac{Z_1 - Z_2}{\text{Variance of } Z \text{ within groups}} // \text{ score function between the classes} \quad (5)$$

For the given score function, the aim is to estimate the linear coefficients of variables that maximize the score, which is further given as:

$$\beta = C^{-1}(\mu_1 - \mu_2) // \text{ coefficients of the model} \quad (6)$$

$$C = \frac{1}{n_1 + n_2} (n_1 C_1 + n_2 C_2) // \text{ Pooled covariance matrix} \quad (7)$$

where,  $\beta$ : coefficients of Linear model,  $C_1$  and  $C_2$ : covariance matrices, and  $\mu_1$  &  $\mu_2$ : mean vectors.

The discriminant assessment can be done by computing the Mahalanobis distance between two groups.

$$Mahalanobis \text{ distance}(\Delta^2) = \beta^T (\mu_1 - \mu_2) \quad (8)$$

At last, we obtain a new data point that can be classified into  $C_1$  (default) and  $C_2$  (not-default) by following the conditional formatting on,

$$\left( x - \left( \frac{\mu_1 + \mu_2}{2} \right) \right) \geq \log \frac{p(C_1)}{p(C_2)} \quad (9)$$

where,  $\beta^T$ : coefficients of vector,  $x$ : vector of the data,  $\frac{\mu_1 + \mu_2}{2}$ : mean value of vector, and  $\frac{p(C_1)}{p(C_2)}$ : probability of class. Depending on the obtained scores, the relevant features are extracted and selected for the classification purpose.

### 2.4. Classification

Iterative support vector machine (ISVM) is employed to ease the classification tasks with minimized computational efforts. It extracts the provider data via feedback loops in an iterative manner. Initially, a hyperplane data cube is created by combining source data tables and their principal components. Then, a general SVM is applied to the hyperplane data cube that generates the classification map. MapReduce framework is employed to receive the required information from the SVM based classification map. The output obtained from the applied preprocessing technique is combined with the other hyperplane data cube for the next iteration process. Likewise, the iterative process continues until achieving the stopping criteria. The proposed steps of the ISVM are given as:

- Initially, let us consider a  $K$  class of interest,  $\{C_p\}_{p=1}^K$ .
- Initializing the conditions as,  $K$  be the number of classes and  $k=1$  and  $\Omega^{(0)} = \{Datatables\} \cup \{PC_1\}$ , where,  $PC_1$  is the principal component of considered data tables.
- Deriving the classification map  $Class - Map_{SVM}^{(0)}$  for the executed SVM on  $\Omega^{(0)}$ .

d) Based on the generated class-map, the  $K$  classification maps are created for  $j^{th}$  iteration. Data  $(x,y)$  of data table under the  $k^{th}$  classification map is represented as,

$$B_{SVM}^{(j)}(x,y) = \{1|if data(x,y) \in class_{SVM}\}; \{0|otherwise\} \tag{10}$$

e) Then, applying the (preprocessing technique) on the  $B_{SVM}^{(j)}(x,y)$  and then filtered the inputs are represented as,  $preprocessing_{SVM}^{(j)}(x,y)$

f) Creating the new hyperplane data cube as,

$$\Omega^{(j)} = \Omega^{(j-1)} \cup \{preprocessing_{SVM,1}^{(j)}\} \cup \{preprocessing_{SVM,n}^{(j)}\}$$

g) Executing the SVM on  $\Omega^{(j)}$  to generate  $Class - Map_{SVM}^{(j)}$

h) Stopping rule is defined for terminating the iteration i.e. Feedback process, which is explained in the next section.

i) If  $Class - Map_{SVM}^{(j)}$ , satisfies the stopping rule, then the ISVM is stopped. Atlast, the final classification map is declared.

j) Else, the process continues by following the step (d), by iteratively,  $j = j+1$ .

### 2.5. Framing of stopping rule for ISVM

The main concept behind the stopping rule of ISVM is to find the best classification maps obtained from  $j^{th}$  and  $(j-1)^{th}$  iterations. Tanimoto index (TI) is employed to find the best stopping rules from the generated classification maps. It is given as,

$$TI^{(j)} = \frac{|Size^j \cap size^{j-1}|}{|Size^j \cup size^{j-1}|} \tag{11}$$

where,  $Size^j$ ;  $size^{j-1}$  are the classification maps.

TI ranges from [0, 1] and a threshold value  $\beta$  is defined. If obtained classification maps cross higher than the given threshold  $\beta$ , then the iteration is stopped. Figure 5 represents the functionality of the ISVM.

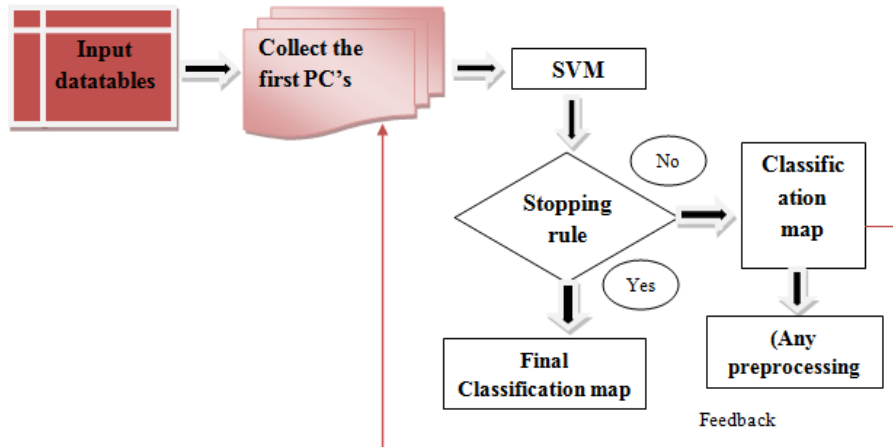


Figure 5. Functional block diagram of the ISVM

### 3. RESULTS AND DISCUSSION

The proposed framework is applied to the real-world insurance provider data obtained from medical fraud provider detection the previous model is compared with the proposed model using institution-level variables. From the Medicare data warehouse, beneficiary, inpatient and outpatient data details are preserved in different tables. The Table 1 presents the tables and their details. It is to be noted that the unique feature between inpatient and outpatient data is the absence of diagnosis code.

Table 1. Used database tables and their details

Data Tables	Details	No. of features
Beneficiary	Basic information of the patients i.e. outpatients as well as inpatients details such as Gender, Claiming details, and Reimbursement details	25
Outpatient	It contains details of the patients who visited hospitals but were not admitted. E.g. <i>ClaimID</i> , <i>ProviderID</i> , and <i>PhysicianID</i>	27
Inpatient	It contains details of the patients admitted in the hospital. E.g. <i>ClaimID</i> , <i>ProviderID</i> , <i>PhysicianID</i> , and <i>Diagnosis code</i>	30

The Table 1 presents the database tables and their feature details. The data is preprocessed using MapReduce technique that eliminates the medical treatment records, claiming records, removing missing values, and fixing errors. As a result, we have used 5,000 records for modeling. Provider ID and the Beneficiary ID are the primary key and the claim details like reimbursement and the deductible amount are taken as the secondary key value of this study. The collected dataset is preprocessed using MapReduce framework. A beneficiary can hold the inpatient and outpatient data and thus, it is organized using MapReduce framework which is given as:

BENE11001 → (Inpatient, 3) & (Outpatient, 0)  
 BENE11002 → (Inpatient, 0) & (Outpatient, 1)  
 BENE11014 → (Inpatient, 1) & (Outpatient, 1)

Table 2 presents the sample records organized using the MapReduce framework. The primary key is to recognize the “clean and organized” data that can reuse the previous results, i.e., it splits the input data into smaller volumes of data quickly and stably. These smaller data volumes may ensure that more small data volumes are clean. Regardless, much smaller data volume increases the overhead, and thus, the designed MapReduce framework, as a preprocessor, must assure stability and speed. In the view of sorting the data imbalance issue, the MapReduce framework adoption has scrutinized the cardinality of the majority and minority classes. Compared to the synthetic minority oversampling technique (SMOTE), the proposed MapReduce technique has modified the intrinsic way of data learning process. The developed Java-based decision support engine is associated with MySQL using java server pages (JSP) scripts. The feature extraction process on the preprocessed data involves claims cost validation. A new data table, ‘Unbundled date’ is created and linked with the proposed (ISVM) classifier. The claims are split into two, namely: i) claims with the approved costs within each diagnostic related group and ii) claims with the disapproved costs within each diagnostic related group.

Table 2. Sample records organized using MapReduce framework

Beneficiary ID	Count
BENE11002	1
BENE11003	2
BENE11004	12
BENE11005	8
BENE11006	1
BENE11007	4
BENE11008	1
BENE11009	2
BENE11132	16
BENE11012	15
BENE11016	15
BENE11024	11
BENE11045	11

LDA is used to haul out a nominal attributes subset that aims for the probability distribution of data classes. The separated classes are close to the original class data distribution by making use of attributes. A new data table is constructed to the estimated ‘confidence score’ of the bills given by the provider. The choice of features based on the LDA are, attendance data, hospital code, diagnostic related group, Claim bill, and drug bill. The dataset is subjected to the ISVM by 70% training and 30% for testing. The approved claims are then fed into ISVM training classifier. The best data those that meet the confidence score of LDA’s criteria are classified first. Each instance of this dataset is organized into “Fraud provider” (or) “Legal Provider”. The proposed iterative conditions fed into the ISVM classifier to detect the fraud providers are: i) count of total *BeneID* is compared with the total *ClaimID* for each provider. If the count of *BeneID* is greater than the count of *ClaimID*, it is labeled as a fraud provider; ii) *claimStartDate* and *ClaimEndDate* are



matched with *PatientAdmissionDate* and *DischargeDate*; and iii) inpatient claim admit diagnosis code is matched with the outpatient diagnosis code.

After each classification, the confusion matrix is displayed. The matrix is embossed of the count of true legal, true fraud, false legal, and false fraud.

- True legal provider: It includes the count of 'approved costs' correctly classified as "True legal provider" by the ISVM classifier.
- True fraud provider: It includes the count of 'disapproved costs' correctly classified as "True fraud provider" by the ISVM classifier.
- False legal provider: It includes the count of 'disapproved costs' incorrectly classified as "False legal provider," even though they are not, by the ISVM classifier.
- False fraud provider: It includes the count of 'approved costs', which were incorrectly classified as "False fraud provider," even though they are not by the ISVM classifier.

The Figure 6 presents the proposed implementation framework. The performance metrics are employed to evaluate the MR-ISVM classifiers.

- Accuracy: The proportion of recognizing the classes to the proportion of aggregate total data samples. The efficacy of the accuracy metric is achieved on the balanced datasets which is expressed as,

$$Accuracy = \frac{TLP+TFP}{TLP+TFP+FLP+FFP} \quad (12)$$

- Precision: The proportion of true legal providers and true fraud providers to the classified positive data samples. It implies the confidence level of the fraud detection, which is expressed as,

$$Precision = \frac{TLP}{TLP+FLP} \quad (13)$$

- Recall: The proportion of true legal providers to the classified positive samples. It implies efficiency of detection rate, which is expressed as,

$$Recall = \frac{TLP}{TLP+FFP} \quad (14)$$

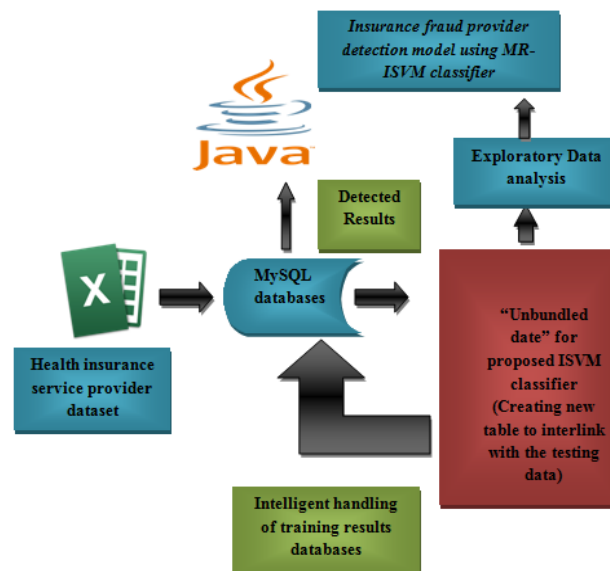


Figure 6. Implementation framework

Table 3 and Figure 7 represents the number of fraud data available in the testing datasets. The accuracy of the MR-ISVM classifier is evaluated from the classification and detection ability of fraudulent providers. The proposed MR-ISVM classifier is tested in 10-fold cross validation of hyperparameters ( $C, \beta$ ). A random search is performed on ISVM parameter training until classifying the optimal claims data samples. The sample screenshots of the proposed framework are shown in Figure 8.

Table 3. Fraud provider types based on data volume

Fraud provider types	Sample data size				
	1,000	2,000	3,000	4,000	5,000
Identity-wise analysis	6	8	45	34	22
Date-wise analysis	0	56	34	12	98
Diagnosis code-wise analysis	45	23	35	122	406

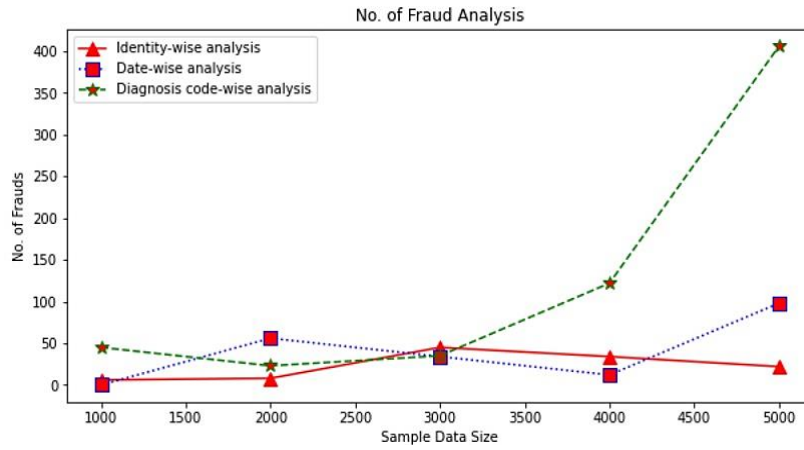


Figure 7. Number of fraud data

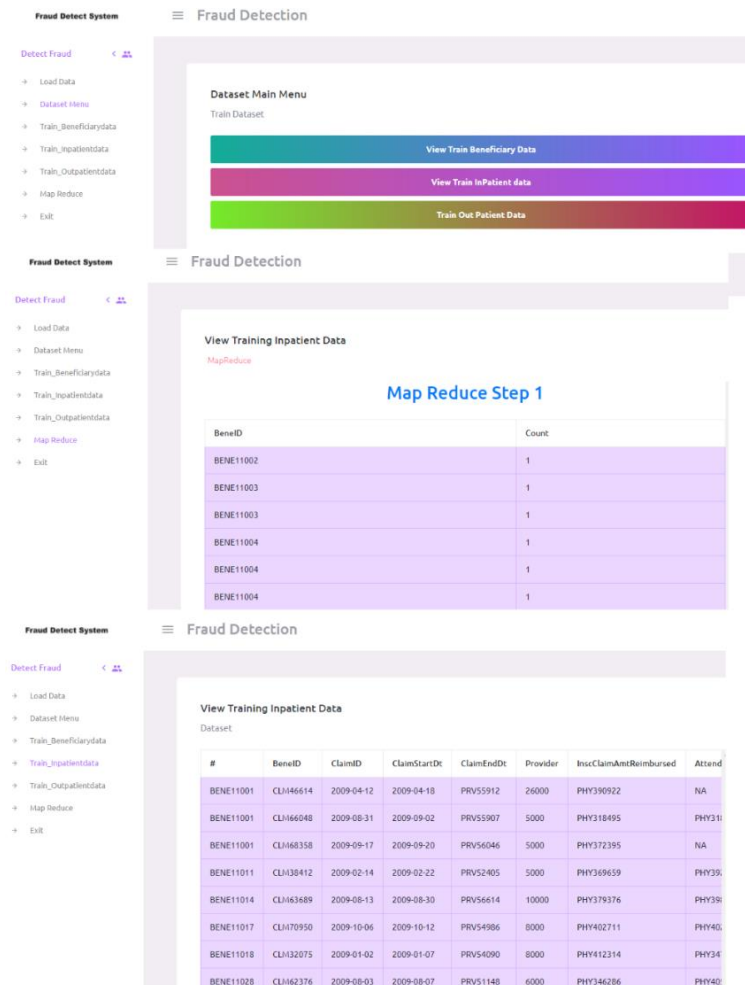


Figure 8. Sample screenshots

From Table 4 and Figures 9 to 11 represents statistics of the SVM classifiers on sample data size. The confusion matrix is also known as the error matrix that helps to visualize the performance of the iterative SVM classifier. As the sample data size increases and given iterative conditions, the classification, and detection of fraud claims incline exponentially.

Table 4. Summary statistics of SVM classifiers on sample sizes

Kernels used	Data size	Accuracy (%)	Precision (%)	Recall (%)
Linear	1000	68.03	56.00	66.90
	2000	73.03	79.80	0.001
	3000	78.90	86.34	65.00
	4000	83.45	78.09	56.12
	5000	73.09	81.03	67.98
Radial basis function	1000	72.45	65.45	63.09
	2000	70.12	87.34	25.67
	3000	81.45	98.33	45.34
	4000	89.34	67.89	39.45
	5000	86.56	85.00	78.90
Iterative loop	1000	73.45	58.91	89.76
	2000	96.78	94.35	98.34
	3000	89.56	96.45	87.46
	4000	83.56	97.88	40.78
	5000	95.34	97.32	83.45

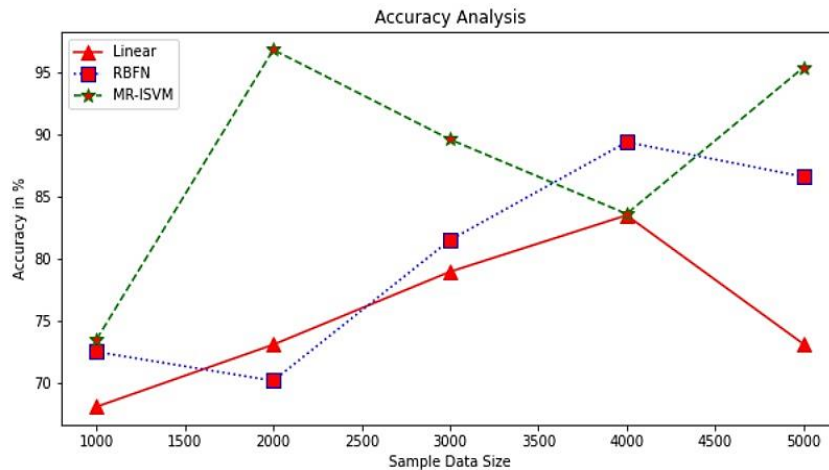


Figure 9. Accuracy analysis

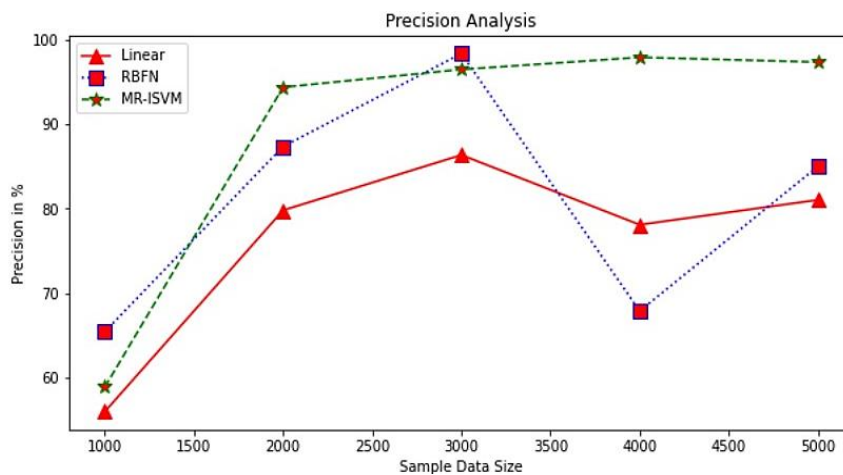


Figure 10. Precision analysis

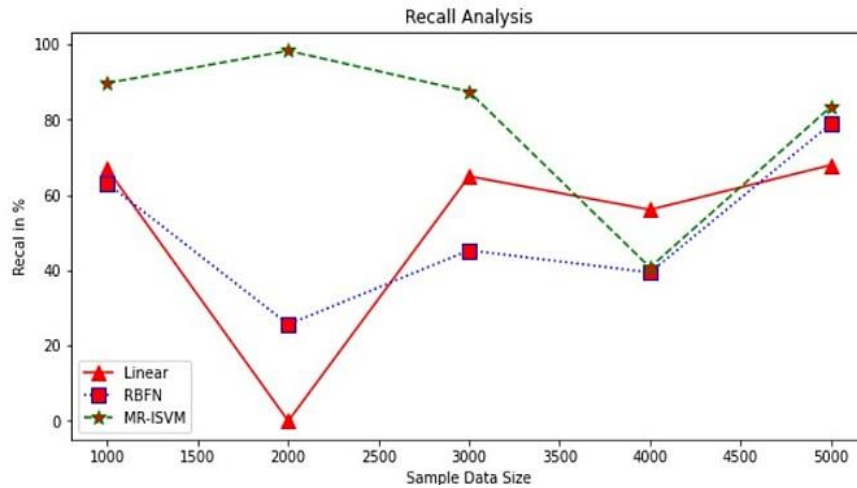


Figure 11. Recall analysis

Table 5 represents the average of different versions of the SVM classifier's performances. It is observed that the MR-ISVM classifiers perform better classification with an accuracy of 87.73%, followed by 88.98% precision and 79.95% recall. Compared to the radial basis function and linear kernels, the MR-ISVM outperformed better to classify and detect the fraud provider.

Figure 12 represent the analysis of the computational time of the MR-ISVM classifier with the linear and radial basis function. Along with the classification, the required time in computing the sample datasets is significant in this study. It is understood from the above analysis that the computational time increases depending on the volume of the sample dataset.

Figure 13 presents the comparative analysis between the existing and proposed techniques. The proposed MR-ISVM classifier takes less computational time than the linear and radial basis functions. The variation in instant time is owing to the training dataset using the MapReduce framework. As we know that the data has been growing widely and rapidly in recent times. Thus, more computational resources need proper and accurate machine learning approaches.

Table 5. Average performance analysis of SVM classifiers

Kernels used	Accuracy (%)	Precision (%)	Recall (%)
Linear	75.3	76.25	51.2
Radial basis function	79.98	80.8	50.49
Iterative loop	87.73	88.98	79.95

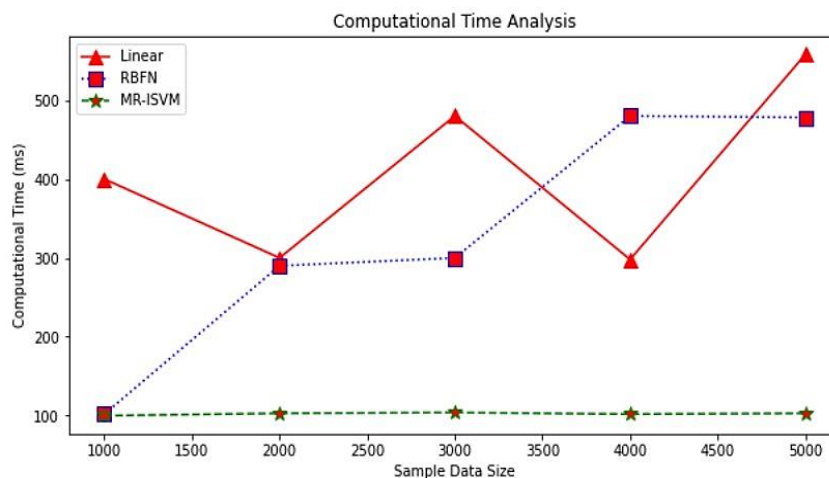


Figure 12. Number computational time analysis

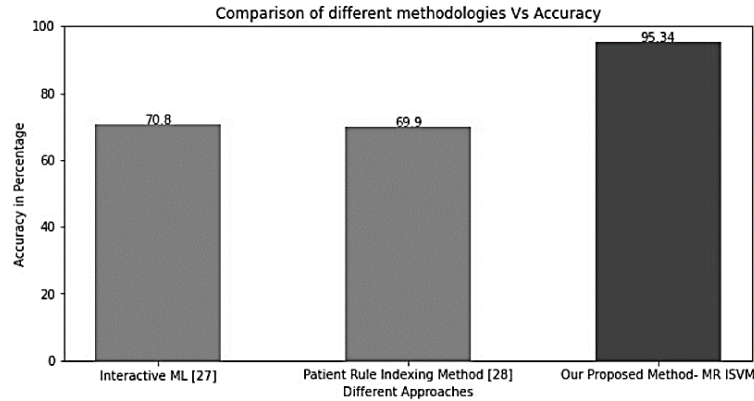


Figure 13. Comparative techniques with proposed MR-ISVM

#### 4. CONCLUSION

The healthcare industry generates a tremendous amount of data from heterogeneous data sources like medical reports, hospital devices, and billing systems. The healthcare data transactions are too complex and voluminous to be computed by conventional methods. Fraud detection is one of the major research areas that need to be scaled up in real-time scenarios. It is a kind of risk management control activity. Class imbalance and feature modeling are the major issues that degrade the performance of machine learning approaches on healthcare data. This research work aims to introduce a novel fraud detection model by hybridizing the qualities of big data and machine learning approaches. The collected insurance claims data is preprocessed using the MapReduce framework that categorizes the voluminous claims data. The required features related to the disease, treatment, and total amount are modeled using LDA. The ISVM approach is widely explored due to its strength in separating the claims data into legal and fraud providers. The soft margin function enables the separation of claims data, which is done by iterative conditions. Thus, the fraud detection systems support the combination of two approaches and achieve higher fraud detection accuracy. The implementation analysis has demonstrated that the MR-ISVM classifier outperforms better in classification and detection than other SVM kernel classifiers. The achieved results explore a positive impact in reducing the computational time on processing healthcare insurance claims without compromising the classification accuracy. The proposed MR-ISVM classifier achieves 87.73% accuracy than the linear (75.3%) and radial basis function (79.98%).





#### REFERENCES

- [1] L. Akoglu, H. Tong, and D. Koutra, "Graph based anomaly detection and description: a survey," *Data Mining and Knowledge Discovery*, vol. 29, no. 3, pp. 626–688, May 2015, doi: 10.1007/s10618-014-0365-y.
- [2] A. A. Aburomman and M. bin Ibne Reaz, "A novel weighted support vector machines multiclass classifier based on differential evolution for intrusion detection systems," *Information Sciences*, vol. 414, pp. 225–246, Nov. 2017, doi: 10.1016/j.ins.2017.06.007.
- [3] I. Sadgali, N. Sael, and F. Benabbou, "Human behavior scoring in credit card fraud detection," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 10, no. 3, pp. 698–706, Sep. 2021, doi: 10.11591/ijai.v10.i3.pp698-706.
- [4] P. P. Vishwakarma, A. K. Tripathy, and S. Vemuru, "An empiric path towards fraud detection and protection for NFC-enabled mobile payment system," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 17, no. 5, pp. 2313–2320, Oct. 2019, doi: 10.12928/telkomnika.v17i5.12290.
- [5] R. A. I. Alhayali, M. Aljanabi, A. H. Ali, M. A. Mohammed, and T. Sutikno, "Optimized machine learning algorithm for intrusion detection," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 24, no. 1, pp. 590–599, Oct. 2021, doi: 10.11591/ijeecs.v24.i1.pp590-599.
- [6] E. Raguseo, "Big data technologies: An empirical investigation on their adoption, benefits and risks for companies," *International Journal of Information Management*, vol. 38, no. 1, pp. 187–195, Feb. 2018, doi: 10.1016/j.ijinfomgt.2017.07.008.
- [7] S. M. Erfani, S. Rajasegarar, S. Karunasekera, and C. Leckie, "High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning," *Pattern Recognition*, vol. 58, pp. 121–134, Oct. 2016, doi: 10.1016/j.patcog.2016.03.028.
- [8] W.-C. Lin, S.-W. Ke, and C.-F. Tsai, "CANN: An intrusion detection system based on combining cluster centers and nearest neighbors," *Knowledge-Based Systems*, vol. 78, pp. 13–21, Apr. 2015, doi: 10.1016/j.knsys.2015.01.009.
- [9] A. Fernández, C. J. Carmona, M. J. del Jesus, and F. Herrera, "A view on fuzzy systems for big data: Progress and opportunities," *Int. Journal of Computational Intelligence Systems*, vol. 9, no. 1, pp. 69–80, 2016, doi: 10.1080/18756891.2016.1180820.
- [10] H. Hassani and E. S. Silva, "Forecasting with big data: A review," *Annals of Data Science*, vol. 2, no. 1, pp. 5–19, Mar. 2015, doi: 10.1007/s40745-015-0029-9.
- [11] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 1153–1176, 2016, doi: 10.1109/COMST.2015.2494502.
- [12] I. A. T. Hashem *et al.*, "The role of big data in smart city," *International Journal of Information Management*, vol. 36, no. 5, pp. 748–758, Oct. 2016, doi: 10.1016/j.ijinfomgt.2016.05.002.





- [13] H. Joudaki *et al.*, "Using data mining to detect health care fraud and abuse: A review of literature," *Global Journal of Health Science*, vol. 7, no. 1, pp. 194–202, Aug. 2014, doi: 10.5539/gjhs.v7n1p194.
- [14] I. Sadgali, N. Sael, and F. Benabbou, "Performance of machine learning techniques in the detection of financial frauds," *Procedia Computer Science*, vol. 148, pp. 45–54, 2019, doi: 10.1016/j.procs.2019.01.007.
- [15] E. M. Hussein Saeed, H. A. Saleh, and E. A. Khalel, "Classification of mammograms based on features extraction techniques using support vector machine," *Computer Science and Information Technologies*, vol. 2, no. 3, pp. 121–131, Nov. 2020, doi: 10.11591/csit.v2i3.p121-131.
- [16] M. D. Salawu *et al.*, "A chi-square-SVM based pedagogical rule extraction method for microarray data analysis," *International Journal of Advances in Applied Sciences*, vol. 9, no. 2, pp. 93–100, Jun. 2020, doi: 10.11591/ijaas.v9.i2.pp93-100.
- [17] M. Moukhafi, K. El Yassini, and B. Seddik, "Intrusions detection using optimized support vector machine," *International Journal of Advances in Applied Sciences*, vol. 9, no. 1, pp. 62–66, Mar. 2020, doi: 10.11591/ijaas.v9.i1.pp62-66.
- [18] J. C. Cassimiro, A. M. Santana, P. S. Neto, and R. L. Rabelo, "Investigating the effects of class imbalance in learning the claim authorization process in the Brazilian health care market," in *2017 International Joint Conference on Neural Networks (IJCNN)*, May 2017, pp. 3265–3272, doi: 10.1109/IJCNN.2017.7966265.
- [19] P. Dora and G. H. Sekharan, "Healthcare in-surance fraud detection leveraging big data analytics," *G. Hari Sekharan*, vol. 4, no. 4, pp. 2073–2076, 2015.
- [20] N. Che and W. Janusz, "Unsupervised labeling of data for supervised learning and its application to medical claims prediction," *Computer Science*, vol. 14, no. 3, pp. 191–214, 2013, doi: 10.7494/csci.2013.14.2.191.
- [21] G. van Capelleveen, M. Poel, R. M. Mueller, D. Thornton, and J. van Hillegersberg, "Outlier detection in healthcare fraud: A case study in the Medicaid dental domain," *International Journal of Accounting Information Systems*, vol. 21, pp. 18–31, Jun. 2016, doi: 10.1016/j.accinf.2016.04.001.
- [22] Y. Gao, C. Sun, R. Li, Q. Li, L. Cui, and B. Gong, "An efficient fraud identification method combining manifold learning and outliers detection in mobile healthcare services," *IEEE Access*, vol. 6, pp. 60059–60068, 2018, doi: 10.1109/ACCESS.2018.2875516.
- [23] L. F. M. Carvalho, C. H. C. Teixeira, W. Meira, M. Ester, O. Carvalho, and M. H. Brandao, "Provider-consumer anomaly detection for healthcare systems," in *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, Aug. 2017, pp. 229–238, doi: 10.1109/ICHI.2017.75.
- [24] D. Thornton, R. M. Mueller, P. Schoutsen, and J. van Hillegersberg, "Predicting healthcare fraud in medicaid: A multidimensional data model and analysis techniques for fraud detection," *Procedia Technology*, vol. 9, pp. 1252–1264, 2013, doi: 10.1016/j.protcy.2013.12.140.
- [25] R. A. Bauder, T. M. Khoshgoftaar, A. Richter, and M. Herland, "Predicting medical provider specialties to detect anomalous insurance claims," in *2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI)*, Nov. 2016, pp. 784–790, doi: 10.1109/ICTAI.2016.0123.
- [26] S. Kareem, R. Binti Ahmad, and A. B. Sarlan, "Framework for the identification of fraudulent health insurance claims using association rule mining," in *2017 IEEE Conference on Big Data and Analytics (ICBDA)*, Nov. 2017, pp. 99–104, doi: 10.1109/ICBDA.2017.8284114.
- [27] I. Kose, M. Gokturk, and K. Kilic, "An interactive machine-learning-based electronic fraud and abuse detection system in healthcare insurance," *Applied Soft Computing*, vol. 36, pp. 283–299, Nov. 2015, doi: 10.1016/j.asoc.2015.07.018.
- [28] S. Sadiq, Y. Tao, Y. Yan, and M.-L. Shyu, "Mining anomalies in Medicare big data using patient rule induction method," in *2017 IEEE Third International Conference on Multimedia Big Data (BigMM)*, Apr. 2017, pp. 185–192, doi: 10.1109/BigMM.2017.56.
- [29] R. Roy and K. T. George, "Detecting insurance claims fraud using machine learning techniques," in *2017 International Conference on Circuit, Power and Computing Technologies (ICCPCT)*, Apr. 2017, pp. 1–6, doi: 10.1109/ICCPCT.2017.8074258.
- [30] C. Sun, Q. Li, H. Li, Y. Shi, S. Zhang, and W. Guo, "Patient cluster divergence healthcare insurance fraudster detection," *IEEE Access*, vol. 7, pp. 14162–14170, 2019, doi: 10.1109/ACCESS.2018.2886680.
- [31] R. A. Gupta, "Medical provider fraud detection dataset," *Kaggle*, 2019. Accessed Oct 15, 2021. [Online]. Available: <https://www.kaggle.com/rohitrox/medical-provider-fraud-detection>.

## BIOGRAPHIES OF AUTHORS



**Jenita Mary Arockiam**     Currently pursuing as a Research Scholar in the Department of Computer Science, College of Science and Humanities, SRM Institute of Science and Technology. Received bachelor's degree from the Computer Science Department, J. A. College affiliated to MK University, Madurai in 2000. Master of Computer Application degree from MGR arts-science College, Periyar University, Salem in 2003. M.Phil. degree from Periyar University, Salem in 2008. Having Twelve years of experience in teaching field and very much interested in big data analytics especially machine learning. She can be contacted at email: ja3368@srmist.edu.in.



**Angelin Claret Seraphim Pushpanathan**     Currently working as an Assistant Professor in the Department of Computer Science at SRM Institute of Science and Technology, Chennai. Received doctorate degree from Bharathiar University, Coimbatore. Having 15 years of teaching experience and published many articles related to component-based power distribution system under various SCI and Scopus indexed journal. Area of research focuses on component-based technology, power distribution system, internet of things, machine learning, artificial intelligence with healthcare. She can be contacted at email: angelins@srmist.edu.in.