Hindawi
Journal of Engineering
Volume 2019, Article ID 1432597, 19 pages
https://doi.org/10.1155/2019/1432597

Research Article

Decision Support System (DSS) for Fraud Detection in Health Insurance Claims Using Genetic Support Vector Machines (GSVMs)

Robert A. Sowah,[1] **Marcellinus Kuuboore**,[1] **Abdul Ofoli**,[2] **Samuel Kwofie**,[3] **Louis Asiedu**,[4] **Koudjo M. Koumadi**,[1] **and Kwaku O. Apeadu**[1]

[1]*Department of Computer Engineering, University of Ghana, PMB 25, Legon, Accra, Ghana*
[2]*Electrical and Computer Engineering Department, University of Tennessee, Chattanooga, TN, USA*
[3]*Department of Biomedical Engineering, University of Ghana, Legon, Accra, Ghana*
[4]*Department of Statistics and Actuarial Science, University of Ghana, Legon, Accra, Ghana*

Correspondence should be addressed to Robert A. Sowah; rasowah@ug.edu.gh

Received 25 January 2019; Accepted 1 August 2019; Published 2 September 2019

Academic Editor: Kamran Iqbal

Fraud in health insurance claims has become a significant problem whose rampant growth has deeply affected the global delivery of health services. In addition to financial losses incurred, patients who genuinely need medical care suffer because service providers are not paid on time as a result of delays in the manual vetting of their claims and are therefore unwilling to continue offering their services. Health insurance claims fraud is committed through service providers, insurance subscribers, and insurance companies. The need for the development of a decision support system (DSS) for accurate, automated claim processing to offset the attendant challenges faced by the National Health Insurance Scheme cannot be overstated. This paper utilized the National Health Insurance Scheme claims dataset obtained from hospitals in Ghana for detecting health insurance fraud and other anomalies. Genetic support vector machines (GSVMs), a novel hybridized data mining and statistical machine learning tool, which provide a set of sophisticated algorithms for the automatic detection of fraudulent claims in these health insurance databases are used. The experimental results have proven that the GSVM possessed better detection and classification performance when applied using SVM kernel classifiers. Three GSVM classifiers were evaluated and their results compared. Experimental results show a significant reduction in computational time on claims processing while increasing classification accuracy via the various SVM classifiers (linear (80.67%), polynomial (81.22%), and radial basis function (RBF) kernel (87.91%).

1. Introduction

Low-income countries have made significant development policy frameworks for the sustainability of growth. These frameworks include healthcare delivery. Ghana is one of the countries which aspired to provide effective and efficient health care. In achieving this noble goal, the National Health Insurance Scheme (NHIS) was established by an Act of Parliament, Act 650, in 2003 [1].

The NHIS, as a social protection initiative, aims at providing financial risk protection against the cost of primary health care for residents of Ghana, and it has replaced the hitherto obnoxious cash and carry system of paying for health care at the point of receiving service. Since its introduction, the scheme has grown to become a significant instrument for financing healthcare delivery in Ghana. For effective and efficient implementation, NHIS introduced a tariff as a standardized primary fee for service rendered to its beneficiaries at their affiliated health institutions. This standardized tool was reviewed in January 2007 by the National Health Insurance Authority (NHIA), the governing body of NHIS, to develop a new tariff for the NHIS due to its expansion of service coverage. The new tariff was developed based on a GDRG (Ghana Diagnostic Related Group) system to include various clinical conditions and surgical procedures grouped under eleven Major Diagnostic

Categories (MDC) or clinical specialties, namely, Adult Medicine, Pediatrics, Adult Surgery, Pediatrics Surgery, Ear, Nose and Throat (ENT), Obstetrics and Gynecology, Dental, Ophthalmology, Orthopedics, Reconstructive Surgery, and Out-Patients' Department (OPD) [2]. These specialties provide a guide to the claim adjudication process and the operational mechanism for reporting claims as well as determine the reimbursement process and create standards of operation between NHIS and service providers [2].

The GDRG code structure uses seven alphanumeric characters. The first four characters represent the MDC or clinical specialty. The next two characters are numbers to represent the number of GDRG within MDC. The last character (A or C) represent the age categories. An "A" represents those greater than or equal to 12 years, and C stands for those less than 12 years.

The World Health Organization (WHO) provided an International Classification of Diseases (ICD-10) to meet the requirements for claim submission [3, 4], but NHIS utilized the GDRG codes since they have full control over them. Hence, the GDRG codes are used to develop the fraud detection model.

A claim is a detailed invoice that service providers send to the health insurer, which shows exactly what services a patient or patients received at the point of healthcare service delivery. Claim processing is the major challenge of providers under the Health Insurance Scheme (HIS) globally due to the excessive fraud in submitted claims and gaming of the system through well-coordinated schemes to siphon money from its coffers [5–9].

Fraud in health care is classified into three categories, namely. (1) service provider (hospitals and physicians) fraud, (2) beneficiary (patients) fraud, and (3) insurer fraud [8, 10–13]. Several types of fraud schemes form the basis of this problem in health insurance programs worldwide. These are (1) billing for services not rendered (identity theft and phantom billing), (2) upcoding of services and items (upcoding), (3) duplicate billing, (4) unbundling of claims (unbundling/creative billing), (5) medically unnecessary services (bill padding), (6) excessive services (bill padding), (7) kickbacks, (8) impersonation, (9) ganging, (10) illegal cash exchange for prescription, (11) frivolous use of service, (12) insurance carriers' fraud, (13) falsifying reimbursement, (14) upcoding of service, and (14) insurance subscribers' fraud, among others [9, 13–19].

It was estimated conservatively that at least 3%, or more than $60 billion, of the US's annual healthcare expenditure was lost due to fraud. Other estimates by government and law enforcement agencies placed this loss as high as 10% or $170 billion [9, 12]. In addition to financial loss, fraud also severely hinders the US healthcare system from providing quality care to legitimate beneficiaries [9]. Hence, effective fraud detection is essential for improving the quality and reducing the cost of healthcare services.

The National Health Care Antifraud Association report in [12, 20] intimated that healthcare fraud strips nearly $70 billion from the healthcare industry each year. In response to these realities, the Health Insurance Portability and Accountability Act of 1996 (HIPAA) specifically established healthcare fraud as a federal criminal offense, with the primary crime carrying a federal prison term of up to 10 years in addition to significant financial penalties [8, 21].

This paper presents the hybridized approach of combining genetic algorithms and support vector machines (GSVMs) to solve the health insurance claim classification problem and eliminate fraudulent claims while minimizing conversion and labour costs through automated claim processing. The significant contributions of this paper are as follows: (1) analysis of existing data mining and machine learning techniques (decision tree, Bayesian networks, Naïve–Bayes classifier, and support vector machines) for fraud detection; (2) development of a novel fraud detection model for insurance claims processing based on genetic support vector machines; (3) design, development, and deployment of a decision support system (DSS) which incorporates the fraudulent claim detection model, business intelligence, and knowledge representation for claims processing at NHIS Ghana; (4) development of a user-friendly graphical user interface (GUI) for the intelligent fraud detection system; and (5) evaluation of the health insurance claims fraud detection system using Ghana National Health Insurances Subscribers' data from different hospitals.

The outline of the paper is as follows: Section 1 presents the introduction and problem statement with research objectives. Section 2 outlines the systematic literature review on various machine learning and data mining techniques for health insurance claims fraud detection. Section 3 gives the theoretical and mathematical foundations of genetic algorithms (GA), support vector machines (SVMs), and the hybrid genetic support vector machines (GSVMs) in combating this global phenomenon. Section 4 provides the proposed methodology for the GSVM fraud detection system, its design processes, and development. Section 5 comprises the design and implementation of the genetic support vector machines, while Section 6 presents the key findings of the research with conclusions and recommendations for future work.

2. Literature Review

Researching into health insurance claims fraud domain requires a clear distinctive view on what fraud is because it is sometimes lumped together with abuse and waste. However, fraud and abuse refer to a situation where healthcare service is paid for but not provided or reimbursement of funds is made to third-party insurance companies. Fraud and abuse are further explained as healthcare providers receiving kickbacks, patients seeking treatments that are potentially harmful to them (such as seeking drugs to satisfy addictions), and the prescription of services known to be unnecessary [12, 17–19]. Health insurance fraud is an intentional act of deceiving, concealing, or misrepresenting information that results in healthcare benefits being paid to an individual or group.

Health insurance fraud detection involves account auditing and detective investigation. Careful account auditing can reveal suspicious providers and policyholders. Ideally, it is best to audit all claims one-by-one. However,

auditing all claims is not feasible by any practical means. Furthermore, it is challenging to audit providers without concrete smoking clues. A practical approach is to develop shortlists for scrutiny and perform auditing on providers and patients in the shortlists. Various analytical techniques can be employed in developing audit shortlists.

The most common fraud detection techniques reported through the literature include the use of machine learning, data mining, AI, and statistical methods. The most cost-saving model using the Naïve–Bayes algorithm was used to create a subsample of 20 claims consisting of 400 objects where 50% of objects were classified as fraud and the other 50% classified as legal, which eventually does not give a clear picture of the decision if compared to other classifiers [22].

The integration of multiple traditional methods has emerged as a new research area in combating fraud. This approach could be supervised, unsupervised, or both, for one method to depend on the other for classification. One method may be used as a preprocessing step to modify the data in preparation for classification [9, 23, 24], or at a lower level, the individual steps of the algorithms can be intertwined to create something fundamentally original. Hybrid methods can be used to tailor solutions to a particular problem domain. Different aspects of performance can be specifically targeted, including classification ability, ease of use, and computational efficiency [14].

Fuzzy logic was combined with neural networks to assess and automatically classify medical claims [14]. The concept of data warehousing for data mining purposes in health care was applied to develop an electronic fraud detection application to review service providers on behavioral heuristics and compared to similar service providers. Australia's Health Insurance Commission has explored the online discounting learning algorithm to identify rare cases in pathology insurance data [10, 25–27].

Researchers in Taiwan developed a detection model based on process mining that systematically identified practices derived from clinical pathways to detect fraudulent claims [8].

Results published in [28, 29] used Benford's Law Distributions to detect anomalies in claims reimbursements in Canada. Despite the detection of some anomalies and irregularities, the ability to identify suspected claims is very limited to health insurance claim fraud detection since it applies to service providers with payer-fixed prices.

Neural networks were used to develop an application for detecting medical abuse and fraud for a private health insurance scheme in Chile [30]. The ability to process claims on real-time basis accounts for the innovative nature of this method. The application of association rule mining to examine billing patterns within a particular specialist group to detect these suspicious claims and potential fraudulent individuals was incorporated in [9, 22, 30].

3. Mathematical Foundations for Genetic Support Vector Machines

In the 1960s, John Holland invented genetic algorithms by involving a simulation of Darwinian survival of the fittest as well as the processes of crossover, mutation, and inversion that occurs in other genetics. Holland's inversion demonstrated that, under certain assumptions, GA indeed achieves an optimal balance [31–34]. In contrast with evolution strategies and evolutionary programming, Holland's original goal was not to design algorithms to solve specific problems but rather to formally study the phenomenon of adaptation as it occurs in nature and to develop ways in which the mechanisms of natural adaptation might be imported into computer systems. Moreover, Holland was the first to attempt to put computational evolution on a firm theoretical footing [35].

Genetic algorithms operate through three main operators, namely, (1) reproduction, (2) crossover, and (3) mutation. A typical genetic algorithm requires (1) a genetic representation of the solution domain and (2) a fitness function to evaluate the solution domain [31–34].

Reproduction is controlled by crossover and mutation operators. Crossover is the process whereby genes are selected from the parent chromosomes, and new offsprings are produced. A mutation is designed to add diversity to the population and ensure the possibility of exploring the entire search space. It replaces the values of some randomly selected genes of a chromosome by some arbitrary new values [33, 35].

During the reproduction stage, an individual is assigned a fitness value derived from its raw performance measure given by the objective function.

Support vector machine (SVM) as a statistical machine learning theory was introduced in 1995 by Vapnik and Corte as an alternative technique for polynomial, radial function, and multilayer perceptron classifiers, in which the weights of the neurons are found by solving quadratic programming (QP) problem with linearity, inequality, and equality constraints rather than by solving a nonconvex, unconstrained minimization problem [36–39]. As a novel machine learning technique for binary classification, regression analysis, face detection, text categorization in bioinformatics and data mining, and outlier detection, SVMs face challenges when the dataset is very large due to the dense nature and memory requirement of the quadratic form of the dataset. However, SVM is an excellent example of supervised learning that tries to maximize the generalization by maximizing the margin and supports nonlinear separation using kernelization [40]. SVM tries to avoid overfitting and underfitting. The margin in SVM denotes the distance from the boundary to the closest data points in the feature space.

Given the claims training dataset correspondingly to $x_n : \mathbb{R}^n \in \mathbb{F}$ in the feature space $\mathbb{F}$, the calculated linear hyperplane dividing them into two labelled classes y_i (fraud and legal) can be mathematically obtained as

$$\omega^T x_i + b = 0, \quad \omega \in \mathbb{R}^n, b \in \mathbb{R}. \tag{1}$$

Assuming the training dataset is correctly classified, as shown in Figure 1.

This means that the SVC computes the hyperplane to maximize the margin separating the classes (legal claims and fraud claims).

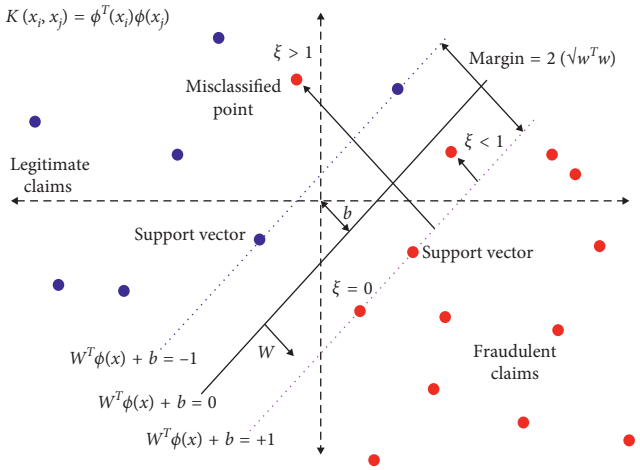

Figure 1: Standard formulation of SVM.

In the simplest linear form, an SVC is a hyperplane that separates the legal claims from the false claims with a maximum margin. Finding this hyperplane involves obtaining two hyperplanes parallel to it, as shown in Figure 1 above, with an equal distance to the maximum margin. If all the training dataset satisfies the constraints as follows:

$$\begin{cases} \omega^T x_i + b \le 1, & \text{for } y_i = +1, \\ \omega^T x_i + b \ge -1, & \text{for } y_i = -1, \end{cases} \tag{2}$$

where ω is the normal to the hyperplane, $|b|/\|\omega\|$ is the perpendicular distance from the hyperplane to the origin, and $\|\omega\|$ is the Euclidean norm of ω. The separating hyperplane is defined by the plane $\omega^T x_i + b = 0$ and the above constraints in (2) are combined to form

$$y_i\left(\omega^T x_i + b\right) \ge 1. \tag{3}$$

The pair of the hyperplanes that gives the maximum margin (γ) can be found by minimizing $\|\omega\|^2$, subject to constraint in (9). This leads to a quadratic optimization problem formulated as

$$\text{Minimize} \quad f(\omega, b) = \frac{\|\omega\|^2}{2}, \tag{4}$$

$$\text{subject to} \quad y_i\left(\omega^T x_i + b\right) \ge 1, \quad \forall i = 1, \ldots, n.$$

This problem is reformulated by introducing Lagrange multipliers, $\alpha_i\,(i = \{1, \ldots, n\})$ for each constraint and subtracting them from the function $f(x) = \omega^T x_i + b$.

This results in establishing the primal Lagrangian function:

$$L_P(\omega, b, \alpha) = \frac{\|\omega\|^2}{2} + \sum_{i=1}^{n}\left(\alpha_i\left(i \ldots y_i\left\{\omega^T x_i + b\right\}\right)\right), \tag{5}$$

$$\forall i = 1, \ldots, n.$$

Taking the partial derivatives of $L_P(\omega, b, \alpha)$ with respect to $\omega, b \,\&\, \alpha$, respectively, and applying the duality theory yields

$$\frac{\partial L_P}{\partial \omega} = 0 \Longrightarrow \omega = \sum_{i=1}^{n} \alpha_i y_i x_i,$$
$$\frac{\partial L_P}{\partial b} = 0 \Longrightarrow b = \sum_{i=1}^{n} \alpha_i y_i. \tag{6}$$

The problem defined in (5) is a quadratic optimization (QP) problem. Maximizing the primal problem L_P with respect to α_i, subject to the constraints that the gradient of L_P with respect to w and b vanish, and that $\alpha_i \ge 0$, gives the following two conditions:

$$\omega = \sum_{i=1}^{n} \alpha_i y_i x_i,$$
$$\sum_{i=1}^{n} \alpha_i y_i = 0. \tag{7}$$

Substituting these constraints gives the dual formulation of the Lagrangian:

$$\underset{\alpha}{\text{Maximize}} \quad L_D(\omega, b, \alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \left(x_i x_j\right),$$

$$\text{subject to} \quad \sum_{i=1}^{n} \alpha_i y_i = 0, \quad \alpha_i \ge 0;\ i = 1, \ldots, n. \tag{8}$$

But the values of α_i, ω, and b are obtained from these respective equations, namely,

$$\omega = \sum_{i=1}^{n} \alpha_i y_i x_i,$$
$$b = \frac{1}{2}\left(Min_{i:y_i} = +1\omega^T x_i + \text{Max}_{i:y_i} = -1\omega^T x_i\right). \tag{9}$$

Also, the Lagrange multiplier is computed using

$$\alpha_i\left(1 - y_i\left(\omega^T x_i + b\right)\right) = 0. \tag{10}$$

Hence, this dual Lagrangian L_D is maximized with respect to its nonnegative α_i to give a standard quadratic optimization problem. The respective training vectors are called support vectors. With the input dataset x_i as a nonzero Lagrangian multiplier α_i,

$$y_i\left(\omega^T x_i + b\right) = 1. \tag{11}$$

The equation above gives the support vectors (SVs).

Despite that the SVM classifier can only have a linear hyperplane as its decision surface, its formulation can be extended to build a nonlinear SVM. SVMs with nonlinear decision surfaces can classify nonlinearly separable data by introducing a soft margin hyperplane, as shown in Figure 2:

Introducing the slack variable into the constraints yields

$$\omega^T x_i + b \ge 1 - \xi_{i,} \quad \text{for } y_i = +1,$$
$$\omega^T x_i + b \ge -1 + \xi_i, \quad \text{for } y_i = -1, \tag{12}$$
$$\xi_{i,} \ge 0, \ \forall i.$$

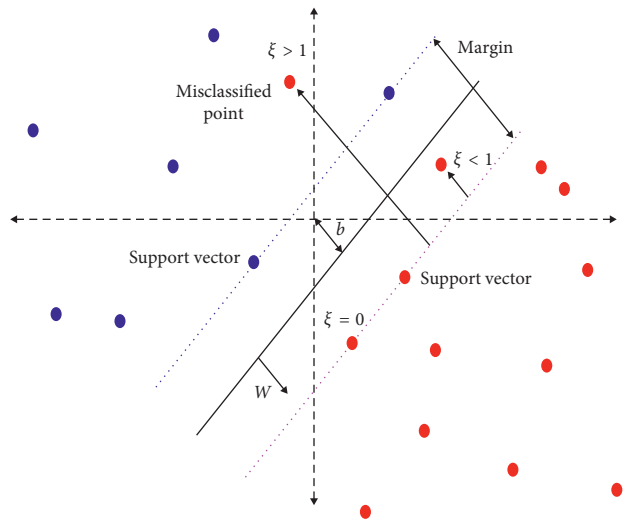

FIGURE 2: Linear separating hyperplanes for the nonseparable case of SVC by introducing the slack variable (ξ).

These slack variables help to find the hyperplane that provides the minimum number of training errors. Modifying equation (4) to include the slack variable yields

$$\underset{\alpha,b,\xi_i}{\text{Minimize}} \quad \frac{\|\omega\|^2}{2} + C\sum_{i=1}^{n}\xi_i,$$

$$\text{subject to} \quad \alpha_i\left(1 - y_i\left(\omega^T x_i + b\right)\right) + \xi_i - 1 \geq 0, \quad \xi_i \geq 0. \tag{13}$$

The parameter C is a regularization parameter that trades off the wide margin with a small number of margin failures. The parameter C is finite. The larger the C value, the more significant the error.

The Karush–Kuhn–Tucker (KKT) conditions are necessary to ensure optimality of the solution to a nonlinear programming problem:

$$\begin{aligned}\left(y_i\left(\omega^T x_i + b\right)\right) - 1 \geq 0, \quad i = 1, 2, \ldots, l, \ \forall i,\\ \alpha_i\left(y_i\left(\omega^T x_i + b\right)\right) - 1 = 0, \quad \alpha_i \geq 0, \ \forall i.\end{aligned} \tag{14}$$

The KKT conditions for the primal problem are used in the nonseparable case, after which the primal Lagrangian becomes

$$\begin{aligned}L_P = \frac{\|\omega\|^2}{2} + c\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{j}\left(\alpha_i\left(y_i\left\{\omega^T x_i + b\right\}\right) - 1 + \xi_i\right)\\ - \sum_{i=1}^{n}\beta_i\xi_i.\end{aligned} \tag{15}$$

With β_i as the Lagrange multipliers to enforce positivity of the slack variables (ξ_i) and applying the KKT conditions to the primal problem yields

$$\begin{aligned}\frac{\partial L_P}{\partial \omega_u} &= \omega_u - \sum_{i=1}^{n}\alpha_i y_i x_{iu} = 0,\\ \frac{\partial L_P}{\partial \omega_u} &= C - \alpha_i y_i = 0,\\ \frac{\partial L_P}{\partial \xi_i} &= C - \alpha_i - \beta_i = 0,\\ \alpha_i\left(y_i\left\{\omega^T x_i + b\right\}\right) - 1 + \xi_i &= 0,\\ y_i\left(\omega^T x_i + b\right) - 1 + \xi_i &\geq 0,\\ \alpha_i, \beta_i, \xi_i \geq 0, \quad C, \xi_i &\geq 0,\\ i = 1, 2, \ldots, n \text{ and } u &= 1, 2, \ldots, d,\\ \frac{\partial L_P}{\partial \omega_u} &= \omega_u - \sum_{i=1}^{n}\alpha_i y_i x_{iu} = 0,\\ \frac{\partial L_P}{\partial \omega_u} &= C - \alpha_i y_i = 0,\end{aligned} \tag{16}$$

where the parameter d represents the dimension of the dataset.

Observing the expressions obtained above after applying KKT conditions yields $\xi_i = 0$ for $\alpha_i < C$, since $\beta_i = C - \alpha_i \neq 0$. This implies that any training point for which $0 < \alpha_i < C$ will be taken to compute for b as a data point that does not cross the boundary:

$$\begin{aligned}\alpha_i &= 0,\\ y_i\left(\omega^T x_i + b\right) - 1 + \xi_i &> 0.\end{aligned} \tag{17}$$

This does not participate in the derivation of the separating function with $\alpha_i = C$ and $\xi_i > 0$:

$$\begin{aligned}\alpha_i &= 0,\\ y_i\left(\omega^T x_i + b\right) - 1 + \xi_i &= 0.\end{aligned} \tag{18}$$

Nonlinear SVM maps the training samples from the input space into a higher-dimensional feature space via a kernel mapping function F. In the dual Lagrangian function, the inner products are replaced by the kernel function:

$$\left(\Phi\left(x_i\right) \cdot \Phi\left(x_j\right)\right) = k\left(x_i, x_j\right). \tag{19}$$

Effective kernels are used in finding the separating hyperplane without high computational resources. The nonlinear SVM dual Lagrangian

$$L_D(\alpha) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j k\left(x_i \cdot x_j\right), \tag{20}$$

subject to

$$\sum_{i=1}^{n}\alpha_i y_i = 0, \quad 0 \leq \alpha_i; \ i = 1, \ldots, n. \tag{21}$$

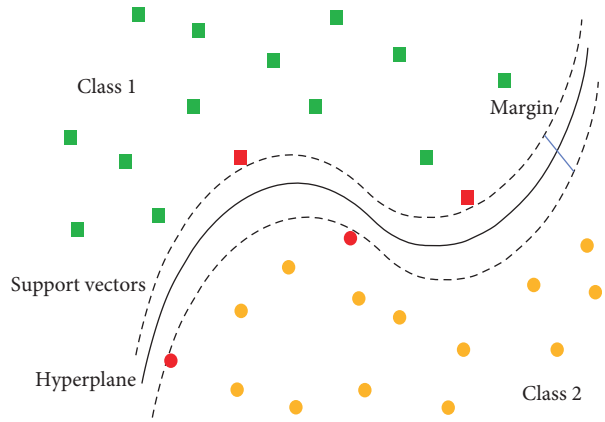

Figure 3: Nonlinear separating hyperplane for the nonseparable case of SVM.

This is like that of the generalized linear case.

The nonlinear SVM separating hyperplane is illustrated in Figure 3 with the support vectors, class labels, and margin.

This model can be solved by the method of optimization in the separable case. Therefore, the optimal hyperplane has the following form:

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i k(x_i, x) + b, \tag{22}$$

where b is the decision boundary from the origin. Hence, separating newly arrived dataset x implies that

$$g(x) = \text{sign}(f(x)). \tag{23}$$

However, feasible kernels must be symmetrical, i.e., the matrix K with the component $k(x_i, x_j)$ is positive semidefinite and satisfies Mercer's condition given in [39, 40]. The summarized kernel functions considered in this work are given in Table 1.

These kernels satisfied Mercer's condition with RBF or Gaussian kernel, which is the widely used kernel function from the literature. The RBF has an advantage of adding a single free parameter $\gamma > 0$, which controls the width of the RBF kernel as $\gamma = 1/2\sigma^2$, where σ^2 is the variance of the resulting Gaussian hypersphere. The linear kernel is given as $k(x_i, x_j) = x_i \cdot x_j$. Consequently, the training of SVMs used the solution of the QP optimization problem. The above mathematical formulations form the foundation for the development and deployment of genetic support vector machines as the decision support tool for detecting and classifying health insurance fraudulent claims. In recent times, decision-making activities of knowledge-intensive enterprises depend holistically on the successful classification of data patterns, despite time and computational resources required to achieve the results due to the complexity associated with the dataset and its size.

Table 1: Summarized kernel functions used.

Kernel name	Parameters	Kernel function
Radial basis function (RBF)	$\gamma \in R$	$k(x_i, x_j) = e^{-\gamma \|x_i - x_j\|^2}$
Polynomial function	$c \in R, d \in N$	$k(x_i, x_j) = (x_i \cdot x_j + c)^d$

4. Methodology for GSVM Fraud Detection

The systematic approach adopted for the design and development of genetic support vector machines for health insurance claims fraud detection is presented in the conceptual framework in Figure 4 and the flow chart implementation in Figure 5.

The conceptual framework incorporates the design and development of key algorithms that enable submitted claims data to be analysed and a model to be developed for testing and validation. The flow chart presents the algorithm implemented based on theoretical foundations in incorporating genetic algorithms and support vector machines, two useful machine learning algorithms necessary for fraud detection. Their combined use in the detection process generates accurate results. The methodology for the design and development of genetic support vector machines as presented above consists of three (3) significant steps, namely, (1) data preprocessing, (2) classification engine development, and (3) data postprocessing.

4.1. Data Preprocessing. The data preprocessing is the first significant stage in the development of the fraud detection system. This stage involves the use of data mining techniques to transform the data from its raw form into the required format to be used by the SVC for the detection and identification of health insurance claims fraud.

The data preprocessing stage involves the removal of unwanted customers, missing records, and data smoothening. This is to make sure that only useful and relevant information is extracted for the next process.

Before the preprocessing, the data were imported from MS Excel CSV format into MySQL to a created database called NHIS. The imported data include the electronic Health Insurance Claims (e-HIC) data and the HIC tariff datasets as tables imported into the NHIS. The e-HIC data preprocessing involves the following steps: (1) claims data filtering and selection, (2) feature selection and extraction, and (3) feature adjustment.

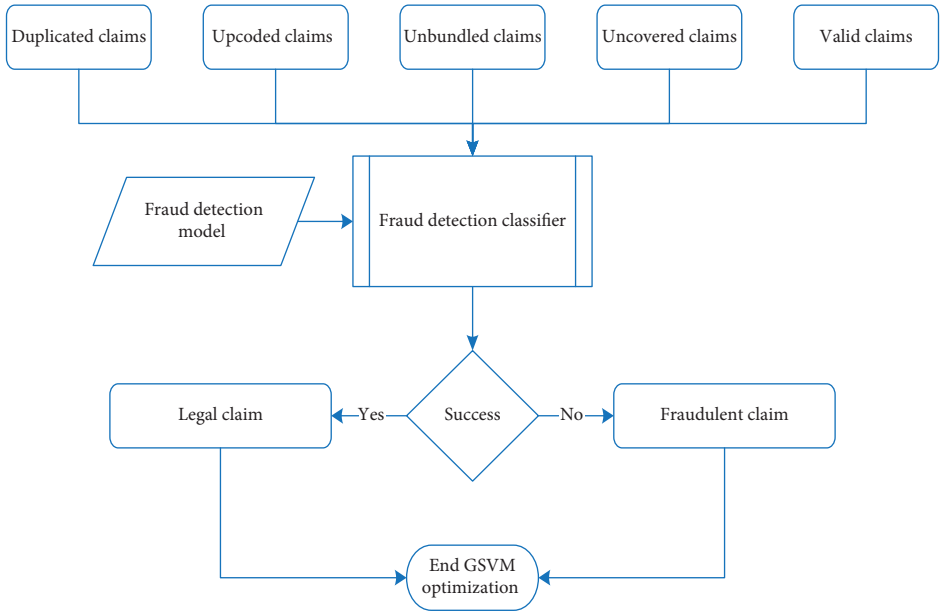

Figure 4: Conceptual model design and development of the genetic support vector machines.

The WEKA machine learning and knowledge analysis environment were used for feature selection and extraction, while the data processing codes are written in the MATLAB technical computing environment. The developed MATLAB-based decision support engine was connected via MYSQL using the script shown in Figure 6.

Preprocessing of the raw data involves claims cost validity checks. The tariff dataset consists of the approved tariffs for each diagnostic-related group, which was strictly enforced to clean the data before further processing. Claims are partitioned into two, namely, (1) claims with the valid and approved cost within each DRG and (2) claims with invalid costs (those above the approved tariffs within each DRG).

With the recent increase in the volume of real dataset and dimensionality of the claims data, there is the urgent need for a faster, more reliable, and cost-effective data mining technique for classification models. The data mining techniques require the extraction of a smaller and optimized set of features that can be obtained by removing largely redundant, irrelevant, and unnecessary features for the class prediction [41].

Feature selection algorithms are utilized to extract a minimal subset of attributes such that the resulting probability distribution of data classes is close to the original distribution obtained using all attributes. Based on the idea of survival of the fittest, a new population is constructed to comply with fittest rules in the current population, as well as the offspring of these rules. Offsprings are generated by applying genetic operators such as crossover and mutation. The process of offspring generation continues until it evolves a population N where every rule in N satisfies the fitness threshold. With an initial population of 20 instances, generation continued till the 20^{th} generation with crossover probability of 0.6 and mutation probability of 0.033. The selected features based on genetic algorithms are "Attendance date," "Hospital code," "GDRG code," "Service bill," and "Drug bill." These are the features selected, extracted, and used as the basis for the optimization problem formulated below:

$$\begin{aligned}
&\text{Minimize} \quad \text{Total}_{\cos t} = f\left(S_{\text{bill}}, D_{\text{bill}}\right),\\
&\text{subject to} \quad \sum_{i=1}^{n}\left(S_{i,bill}\right) \le G_{tariff}, \quad \forall i,\ i = 1, 2, \ldots, n,\\
&\qquad\qquad \sum_{i=1}^{n}\left(D_{i,bill}\right) \le D_{tariff}, \quad \forall j,\ j = 1, 2, \ldots, n,\\
&\qquad\qquad S_{\text{bill}} \text{ is the Service bill},\\
&\qquad\qquad D_{\text{bill}} \text{ is the Drug bill.}
\end{aligned} \tag{24}$$

The GA e-HIC dataset is subjected to SVM training, using 70% of the dataset and 30% for testing as depicted in Figure 7.

The e-HIC dataset, which passes the preprocessing stage, that is the valid claims, was used for SVM training and testing. The best data, those that meet the genetic algorithm's criteria, are classified first. Each record of this dataset is classified as either "*Fraudulent Bills*" or "*Legal Bills*."

The same SVM training and the testing dataset are applied to the SVM algorithm for its performance analysis. The inbuilt MATLAB code for SVM classifiers was integrated as one function for linear, polynomial, and RBF kernels. The claim datasets were partitioned for the classifier training, testing, and validation, 70% of the dataset was used for training, and 30% used for testing. The linear, polynomial, and radial basis function SVM classification kernels were used with ten-fold cross validation for each kernel and the results averaged. For the polynomial classification kernel, a cubic polynomial was used. The RBF classification kernel used the SMO method [40]. This method ensures the

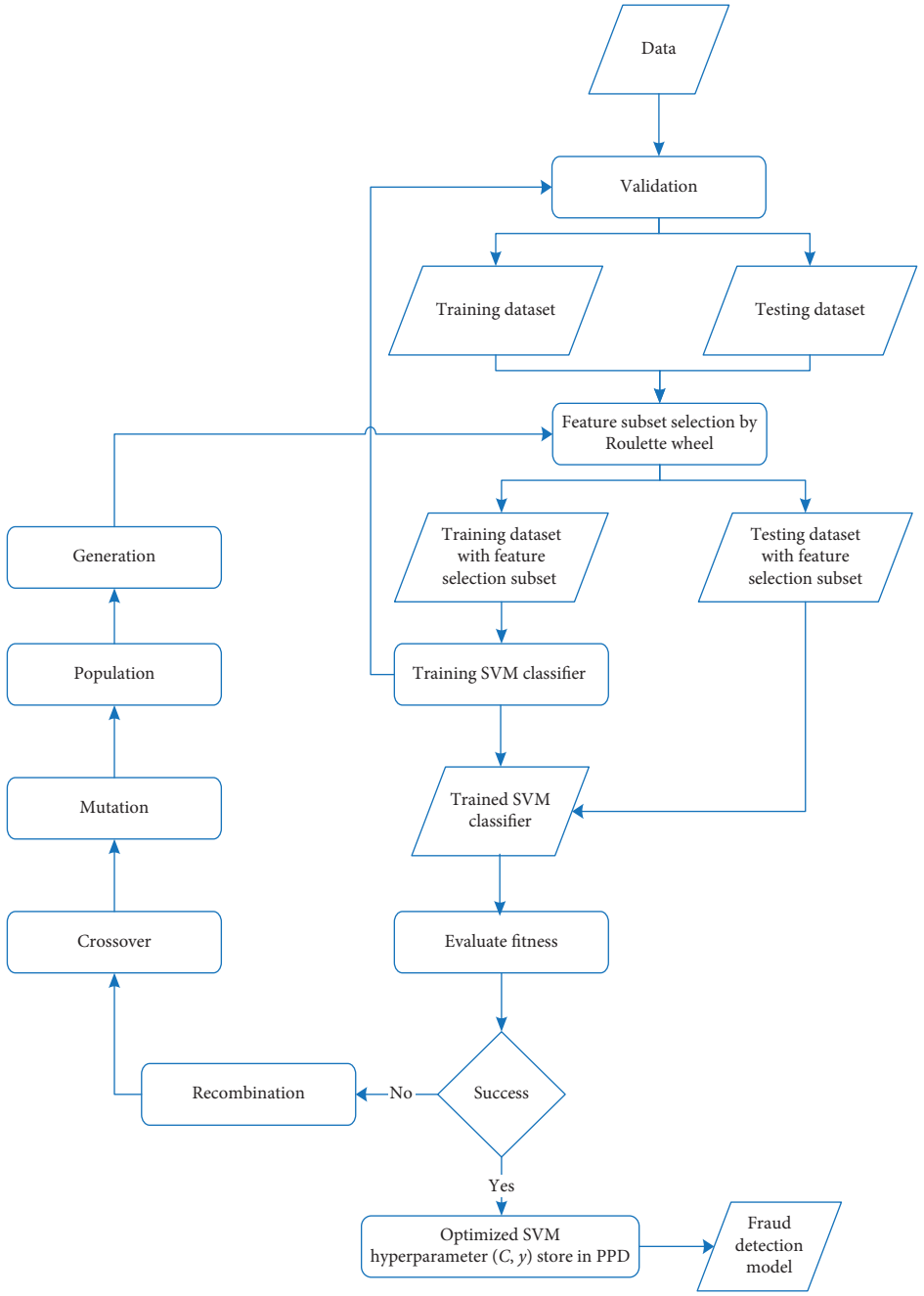

Figure 5: Flow chart for design and development of the genetic support vector machines.

handling of large data sizes as it does data transformation through kernelization. After running many instances and varying parameters for RBF, a variance of 0.9 gave better results as it corresponded well with the datasets for the classification. After each classification, the correct rate is calculated and the confusion matrix extracted. The confusion matrix gives a count for the *true legal*, *true fraudulent*, *false legal, false fraudulent,* and *inconclusive bills*.

(i) True legal bills: this consists of the number of "Legal Bills," which were correctly classified as "Legal Bills" by the classifier.

(ii) True fraudulent bills: this consists of the number of "Fraudulent Bills," which were correctly classified as "Fraudulent Bills" by the classifier.

(iii) False legal bills: this consists of the bills classified as "Legal Bills" even though they are not. That is, these are wrongly classified as "Legal Bills" by the kernel used.

(iv) False, fraudulent bills: the classifier also wrongly classified bills as fraudulent. The confusion matrix gives a count of these wrongly or incorrectly classified bills.

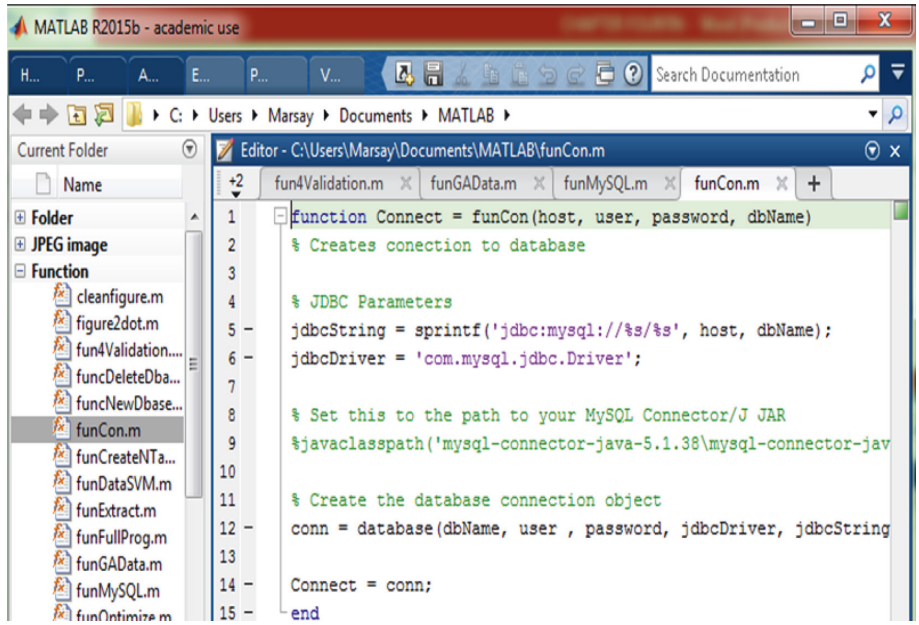

Figure 6: MATLAB-based decision support engine connection to the database.

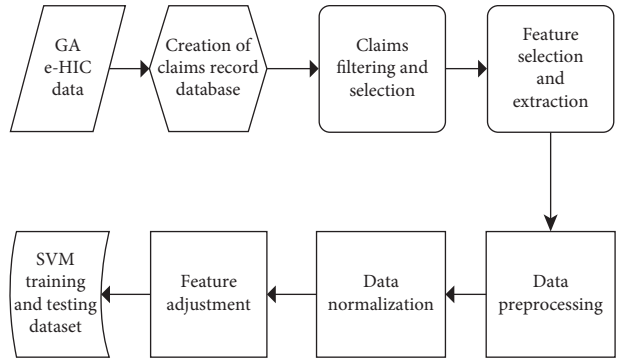

Figure 7: Data preprocessing for SVM training and testing.

(v) Inconclusive bills: these consist of nonclassified bills.

The correct rate: this is calculated as the total number of correctly classified bills, namely, the true legal bills and true fraudulent bills, divided by the total number of bills used for the classification:

$$\text{correct rate} = \frac{\text{number of TLB} + \text{number of TFB}}{\text{total number of bills (TB)}}, \quad (25)$$

where TLB: True Legal Bills; TFB: True Fraudulent Bills.

$$\text{accuracy} = (1 - \text{Error}) = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} = \Pr(C). \quad (26)$$

The probability of a correct classification.

4.1.1. Sensitivity. This is the statistical measure of the proportion of actual fraudulent claims which are correctly detected:

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{\text{TP}}{\text{PP}}. \quad (27)$$

4.1.2. Specificity. This is the statistical measure of the proportion of negative fraudulent claims which are correctly classified:

$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} = \frac{\text{TN}}{\text{NP}}. \quad (28)$$

4.2. GSVM Fraud Detection System Implementation and Testing. The decision support system comprises four main modules integrated together, namely, (1) algorithm implementation using MATLAB technical computing platform, (2) development of graphical user interface (GUI) for the HIC fraud detection system which consists of uploading and processing of claims management, (3) system administrator management, and (4) postprocessing of detection and classification results.

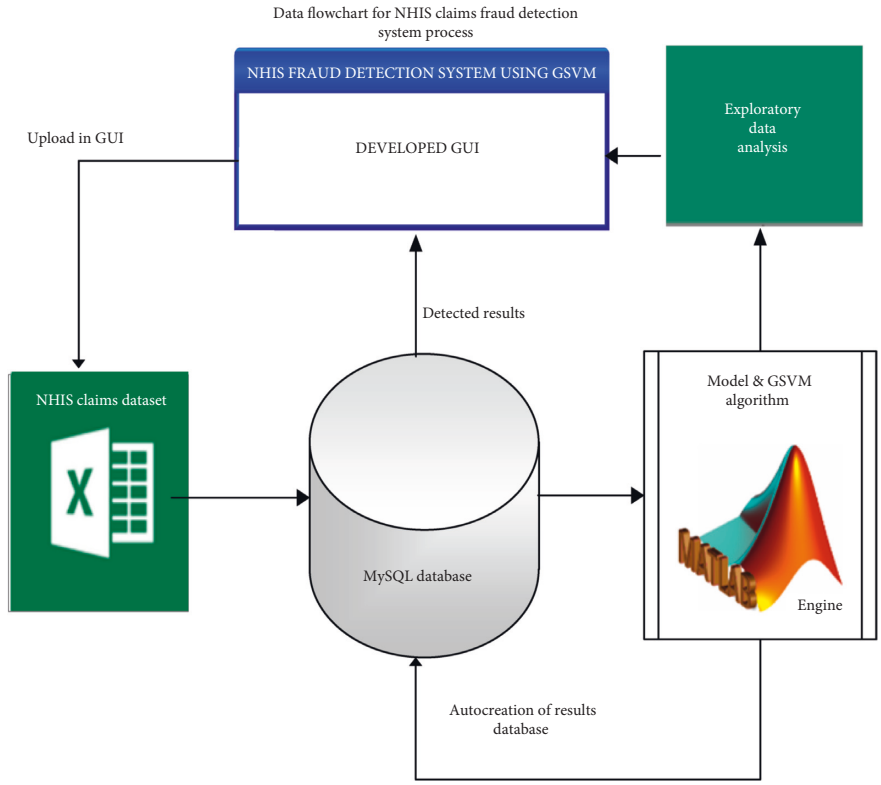

Figure 8: System implementation architecture for HICFDS.

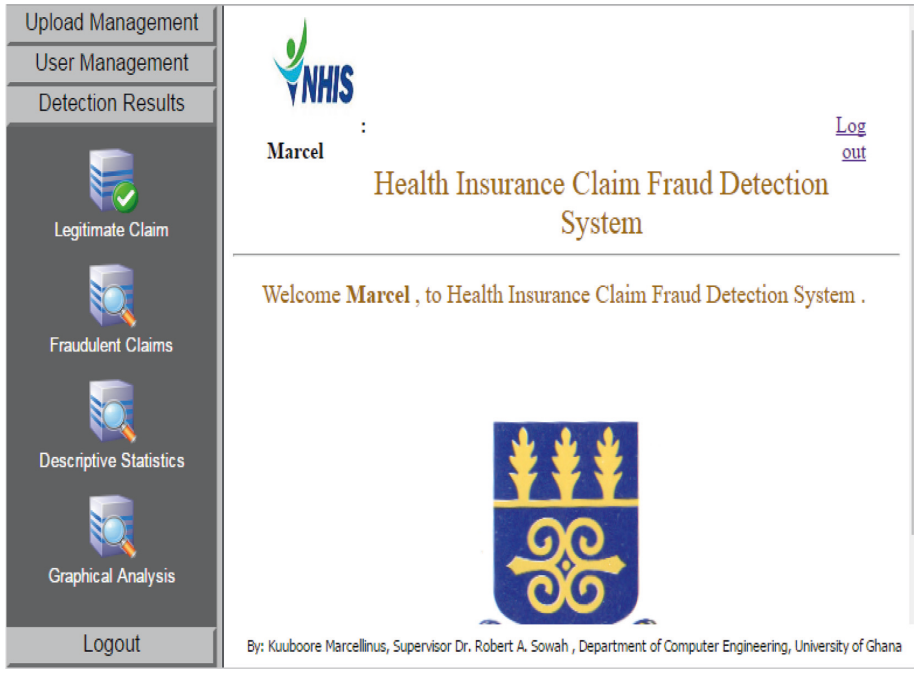

Figure 9: Detection results control portal interface.

The front end of the detection system was developed using XAMPP, a free and open-source cross-platform web server solution stack package developed by Apache Friends [42], consisting mainly of the Apache HTTP Server, MariaDB database, and interpreters for scripts written in the PHP and Perl programming languages [42]. XAMPP stands for Cross-Platform (X), Apache (A), MariaDB (M), PHP (P), and Perl (P). The Health Insurance Claims Fraud Detection System (HICFDS) was developed using MATLAB technical computing environment with the capability to connect to an

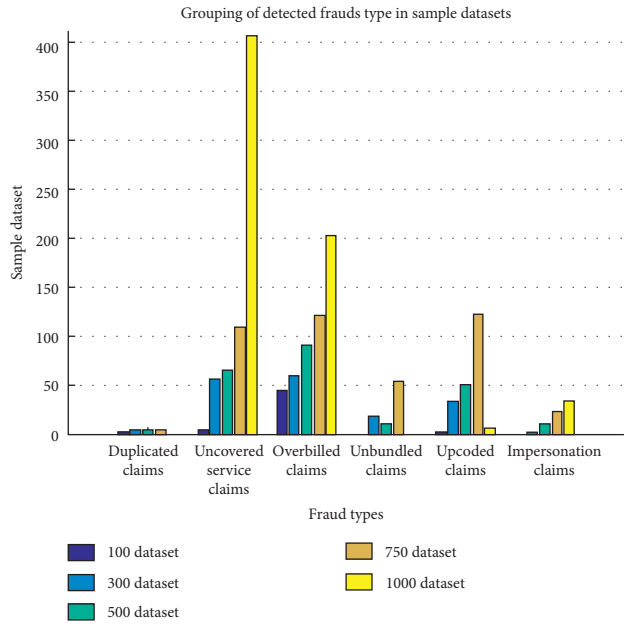

Figure 10: Fraud type distribution on the sample data sizes.

external database and a graphical user interface (GUI) for enhanced interactivity with users. The HICFDS consists of several functional components, namely, (1) function for computing the descriptive statistics of raw and processed data, (2) preprocessing wrapper function for data handling and processing, and (3) MATLAB functions for GA Optimization and SVM Classification processes. The HICFDS components are depicted in Figure 8.

The results generated by the HICFDS are stored in MYSQL database. The results comprise three parts, which are legitimate claims report, fraudulent claims, and statistics of the results. These results are shown in Figure 9. The developed GUI portal for the analysis of results obtained from the classification of the submitted health insurance claims is displayed in Figure 9. By clicking on the fraudulent button in the GUI, a pop-up menu generating the labelled Figure 10 is obtained for the claims dataset. It shows the grouping of detected fraudulent claim types in the datasets.

For each classifier, a 10-fold cross validation (CV) of hyperparameters (C, γ) from Patients Payment Data (PPD) was performed. The performance measured on GA optimization tested several hyperparameters for the optimal SVM. The SVC training aims for the best SVC parameters (C, γ) in building the HICFD classifier model. The developed classifier is evaluated using testing and validation data. The accuracy of the classifier is evaluated using cross validation (CV) to avoid overfitting of SVC during training data. The random search method was used for SVC parameter training, where exponentially growing sequences of hyperparameters (C, γ) as a practical method to identify suitable parameters were used to identify SVC parameters and obtain the best CV accuracy for the classifier claims data samples. Random search slightly varies from grid search. Instead of searching over the entire grid, random search only evaluates a random sample of points on the grid. This makes the random search a computational method cheaper than a grid search. Experimentally, 10-fold CV was used as the measure of the training accuracy, where 70% of each sample was used for training and the remaining 30% used for testing and validation.

Table 2: Sample data size and the corresponding fraud types.

Fraud types	Sample data size				
	100	300	500	750	1000
Duplicate claims	2	4	4	4	0
Uncovered service claims	4	56	65	109	406
Overbilling claims	44	60	91	121	202
Unbundled claims	0	18	10	54	0
Upcoded claims	2	34	50	122	6
Impersonation claims	0	2	10	23	34
Total suspected claims	52	174	230	433	648

Table 3: Summary performance metrics of SVM classifiers on samples sizes.

Kernels used	Data size	Description		
		Average accuracy rate (%)	Sensitivity (%)	Specificity (%)
Linear	100	71.43	60.00	77.78
	300	72.73	84.21	0.00
	500	91.80	97.78	75.00
	750	84.42	95.00	47.06
	1000	82.95	85.42	80.00
Polynomial	100	71.43	66.67	72.73
	300	72.73	88.24	20.00
	500	96.72	100.00	86.67
	750	80.52	96.36	40.91
	1000	84.71	83.67	86.11
Radial basis function	100	71.43	57.14	85.71
	300	95.45	95.00	100.00
	500	99.18	100.00	96.30
	750	82.56	96.88	40.91
	1000	90.91	100.00	82.98

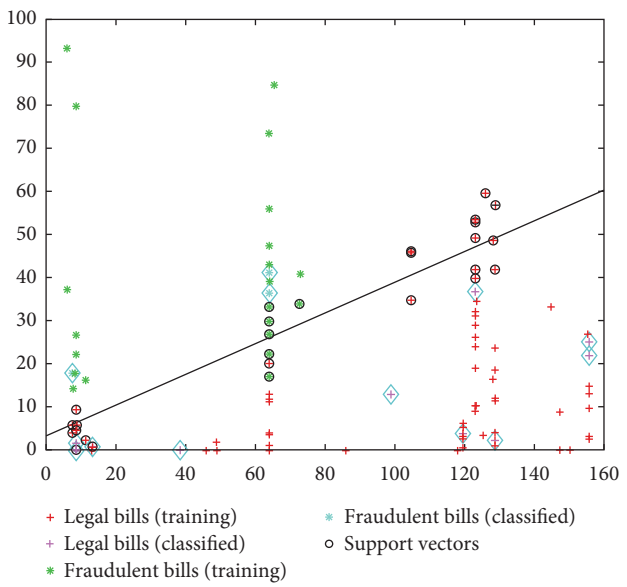

Figure 11: Linear SVM on a sample claims dataset.

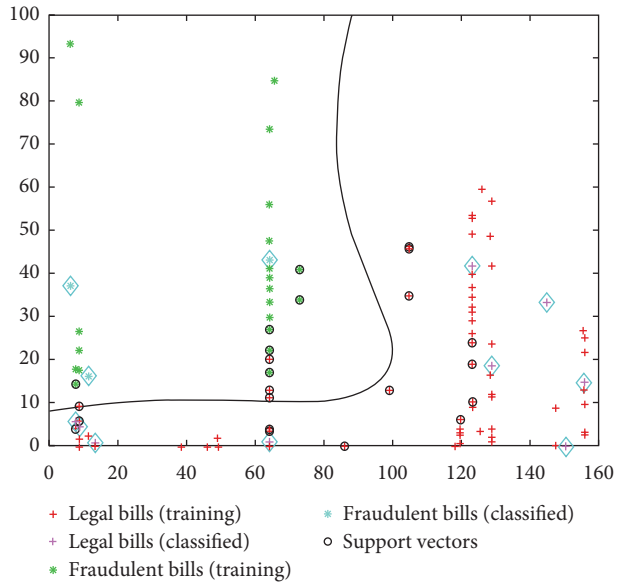

Figure 12: Polynomial SVM on a sample claims dataset.

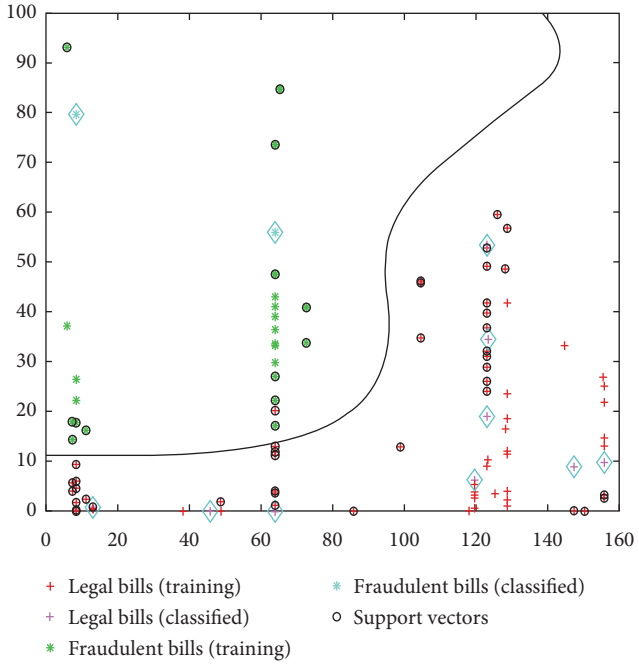

Figure 13: RBF SVM on a sample claims dataset.

Table 4: Averages performance analysis of SVM classifiers.

Description	Accuracy	Sensitivity	Specificity
Linear	80.67	84.48	55.97
Polynomial	81.22	86.99	61.28
RBF	87.91	89.80	81.18

4.3. Data Postprocessing: Validation of Classification Results. The classification accuracy of the testing data is a gauge to evaluate the ability of the HICFDS to detect and identify fraudulent claims. The testing data used to assess and evaluate the efficiency of the proposed HICFDS (classifier) are taken exclusively from NHIS headquarters and covers different hospitals within the Greater Accra Region of Ghana. Sampled data with the corresponding fraud types after the analysis are shown in Table 2.

Table 5: Confusion matrix for SVM classifiers.

Description	Data size	TP	TN	FP	FN	Correct rate
Linear	100	3	7	2	2	71.4
	300	16	0	3	3	71.3
	500	88	24	8	2	91.8
	750	57	8	9	3	84.4
	1000	41	32	8	7	83.0
Polynomial	100	2	8	3	1	71.4
	300	15	1	4	2	72.3
	500	92	26	4	0	96.7
	750	53	91	13	2	80.5
	1000	41	31	5	8	85.2
Radial basis function	100	4	6	1	3	71.4
	300	19	2	0	1	95.5
	500	95	26	1	0	99.2
	750	62	9	13	2	92.2
	1000	41	39	8	0	91.9

In evaluating the classifiers obtained with the analyzed methods, the most widely employed performance measures are used: accuracy, sensitivity, and specificity with their concepts of True Legal (TP), False Fraudulent (FN), False Legal (FP), and True Fraudulent (TN). This classification is shown in Table 3.

The figures below show the SVC plots on the various classifiers (linear, polynomial, and RBF) on the claims datasets (Figures 11–13).

From the performance metrics and overall statistics presented in Table 4, it is observed that the support vector machine performs better classification with an accuracy of 87.91% using the RBF kernel function, followed by the

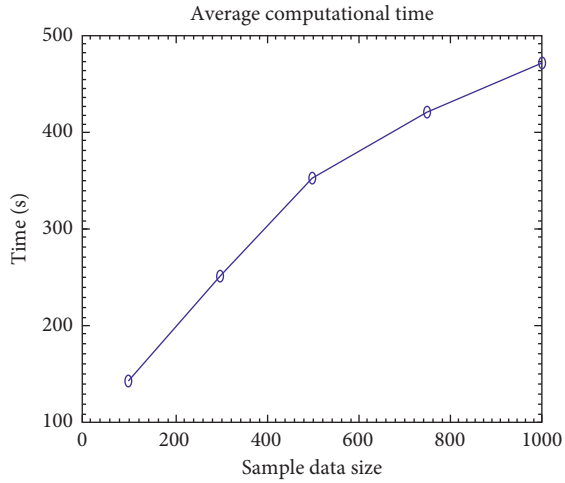

Figure 14: Computational time on the tested sample dataset.

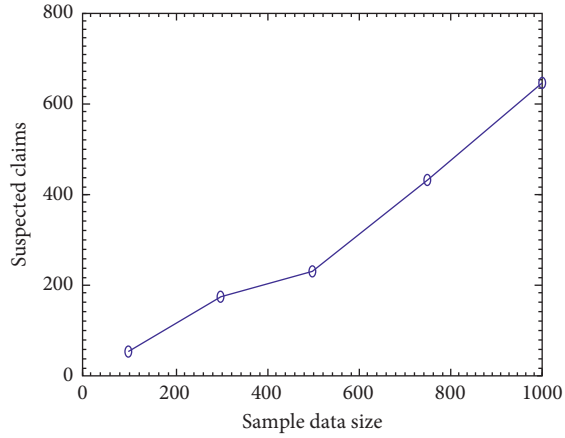

Figure 15: Detected fraud trend on the tested claims dataset.

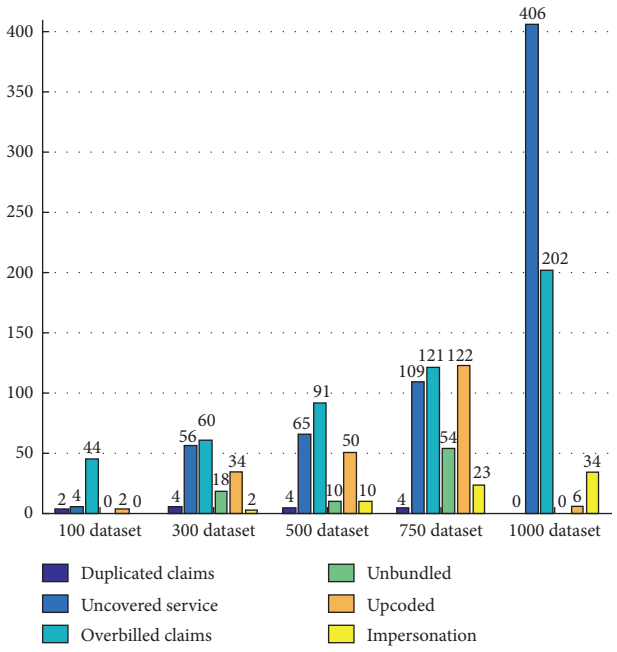

Figure 16: Chart of types of fraudulent claims.

Table 6: Cost analysis of tested claims dataset.

Sample data size	Raw cost of claims (R) GHC	Valid claims cost (V)	Deviation (R–V)	Percentage difference
100	20791.83	8911.72	11880.11	133.31
300	31496.05	15622.7	15873.35	101.60
500	58218.65	27480.96	30737.69	111.85
750	88394.07	31091.58	57302.49	184.30
1000	117448.2	47943.38	69504.82	144.97

polynomial kernel with 81.22% accuracy and hence linear SVM emerging as the least performance classifier with an accuracy of 80.67%. The confusion matrix for the SSVM classifiers is given in Table 5, where i utilized in the computation of the performance metric of the SVM classifiers. For the purpose of statistical and machine learning classification tasks, a confusion matrix, also known as an error matrix, is a specific table layout that allows visualization of the performance of a supervised learning algorithm.

Besides classification, the amount of time required in processing the sample dataset is also an important consideration in this research. From the above, the compared computational time shows that increase in the size of the sample dataset also increases the computational time needed to execute the process regardless of the machine used, which is widely expected. This difference in time costs is merely due to the cause of training the dataset. Thus, as global data warehouse increases, more computational resources will be needed in machine learning and data mining research pertaining to the detection of insurance fraud as depicted in Figure 14 relating the average computational time and sample data.

Figure 15 summarizes the fraudulent claims detected during the testing of the HICFD with the sample dataset used. As the sample data size increases, the number of suspected claims increases rapidly based on the various fraudulent types detected.

Benchmarking HICFD analysis ensures understanding of HIC outcomes. From the chart above, an increase in the claims dataset has a corresponding increase in the number of suspected claims. The graph in Figure 16 shows a sudden rise in the level of *suspected claims* on tested 100 datasets representing 52% of the sample dataset, after which it continues to increase slightly on the suspected numbers of claims by 2% to make up 58% on the tested data size of 300 claims.

Among these fraud types, the most frequent fraudulent act is uncovered services rendered to insurance subscribers by service providers. It accounts for 22% of the fraudulent claims as to the most significant proportion of the total health insurance fraud on the total tested dataset. Consequently, overbilling of submitted claims is recorded as the second fraudulent claims type representing 20% of the total sample dataset used for this research. This is caused by service providers billing for a service greater than the expected tariff to the required diagnoses. Listing and billing for a more complex or higher level of service by providers are done to boost their financial income flow unfairly in the legitimate claims.

Table 7: Comparison of results of GSVM with decision trees and Naïve–Bayes.

Description of the algorithm used	Claims dataset	Accuracy obtained with the corresponding dataset	Average value over different datasets
GSVM with radial basis function (RBF) kernel	100	71.43	87.906
	300	95.45	
	500	99.18	
	750	82.56	
	1000	90.91	
Decision trees	100	62	74.44
	300	78	
	500	77.8	
	750	82.7	
	1000	71.7	
Naïve–Bayes	100	50	59.1
	300	61	
	500	56.8	
	750	60.7	
	1000	67	

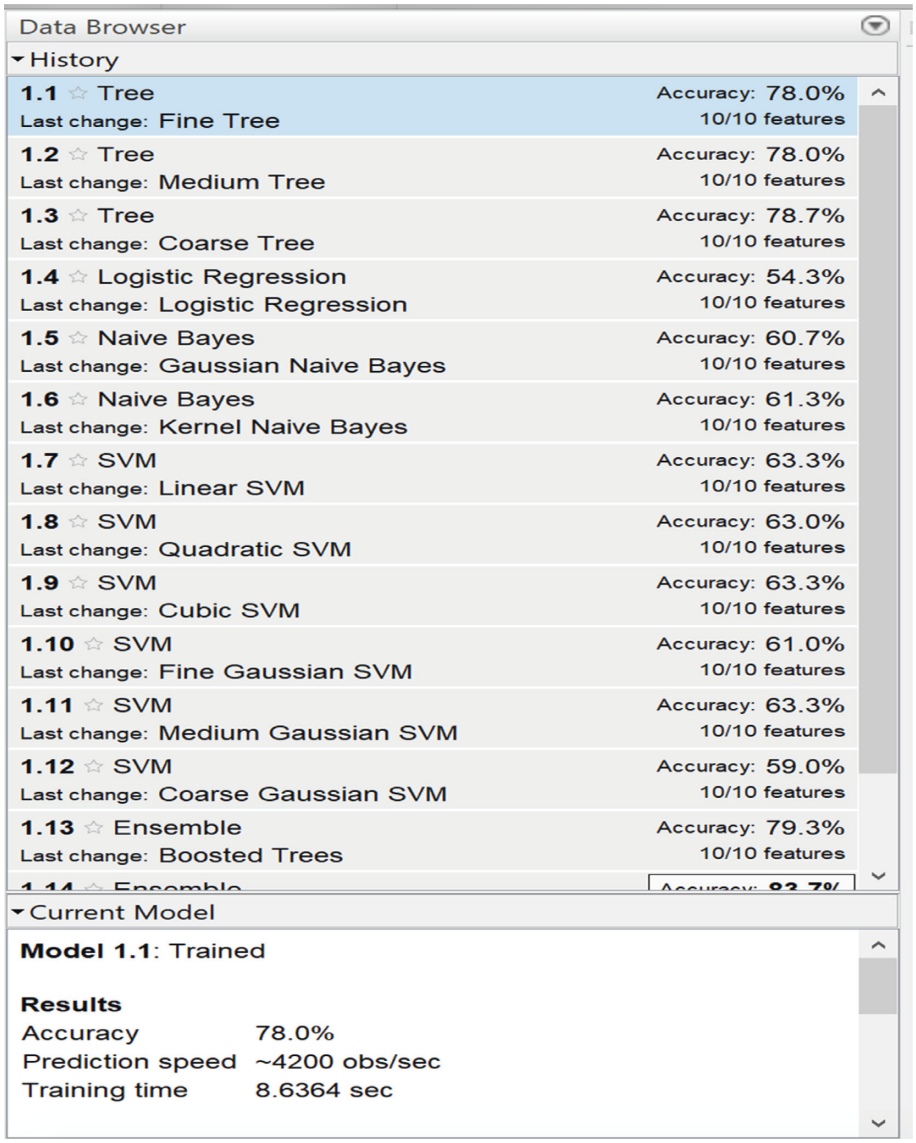

Figure 17: Classification Learner App showing the various algorithms and percentage accuracies in MATLAB.

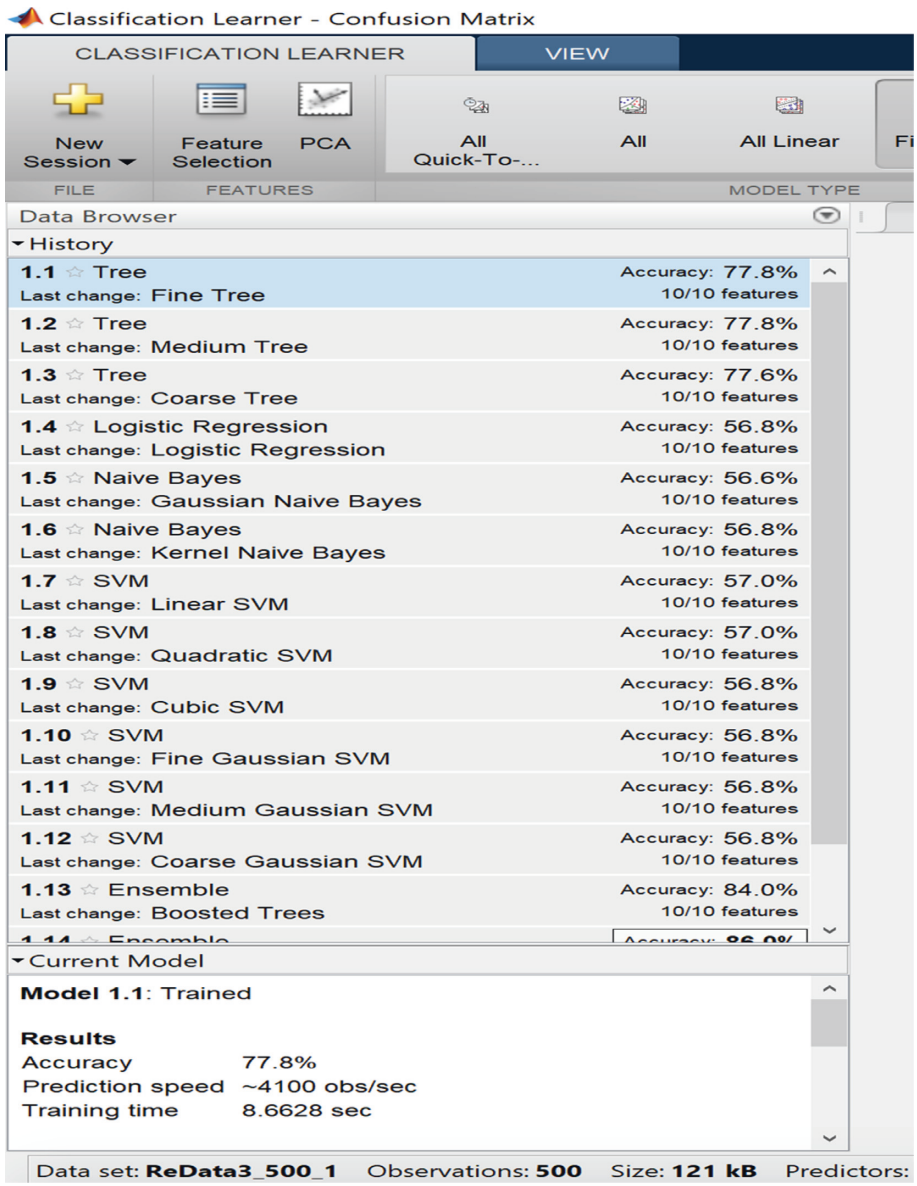

Figure 18: Algorithmic runs on 500-claim dataset.

Moreover, some illicit service providers claim to have rendered service to insurance subscribers on costly services instead of providing more affordable ones. Claims prepared on expensive service rendered to insurance subscribers represent 8% of the fraudulent claims detected on the total sample dataset. Furthermore, 3.1% of service procedure that should be considered an integral part of a single procedure known as the unbundle claims contributed to the fraudulent claims of the set of claims dataset used as the test data. Due to the insecure process for quality delivery of healthcare service, insurance subscribers are also contributing to the fraudulent type of claims by loaning their ID cards to family members of the third party who pretend to be owners and request for the HIS benefits in the healthcare sector. Duplicated claims as part of the fraudulent act recorded the minimum rate of 0.5% of contribution to fraudulent claims in the whole sample dataset.

As observed in Table 6, the cost of the claims bill increases proportionally with an increase in the sample size of the claims bill. This is consistent with an increase in fraudulent claims as sample size increases. From Table 6, we can see the various costs for each raw record (R) of sample claim dataset. Valid claims bill after processing dataset, the variation in the claims bill (R–V), and their percentage representation as well are illustrated in Table 6. There is a 27% financial loss of the total submitted claim bills to insurance carriers. This loss is the highest rate of loss within the 750 datasets of submitted claims.

Summary of results and comparison with other machine learning algorithms such as decision trees and Naïve–Bayes is presented in Table 7.

The MATLAB Classification Learner App [43] was chosen to validate the results obtained above. It enables ease of comparison with the different methods of classification

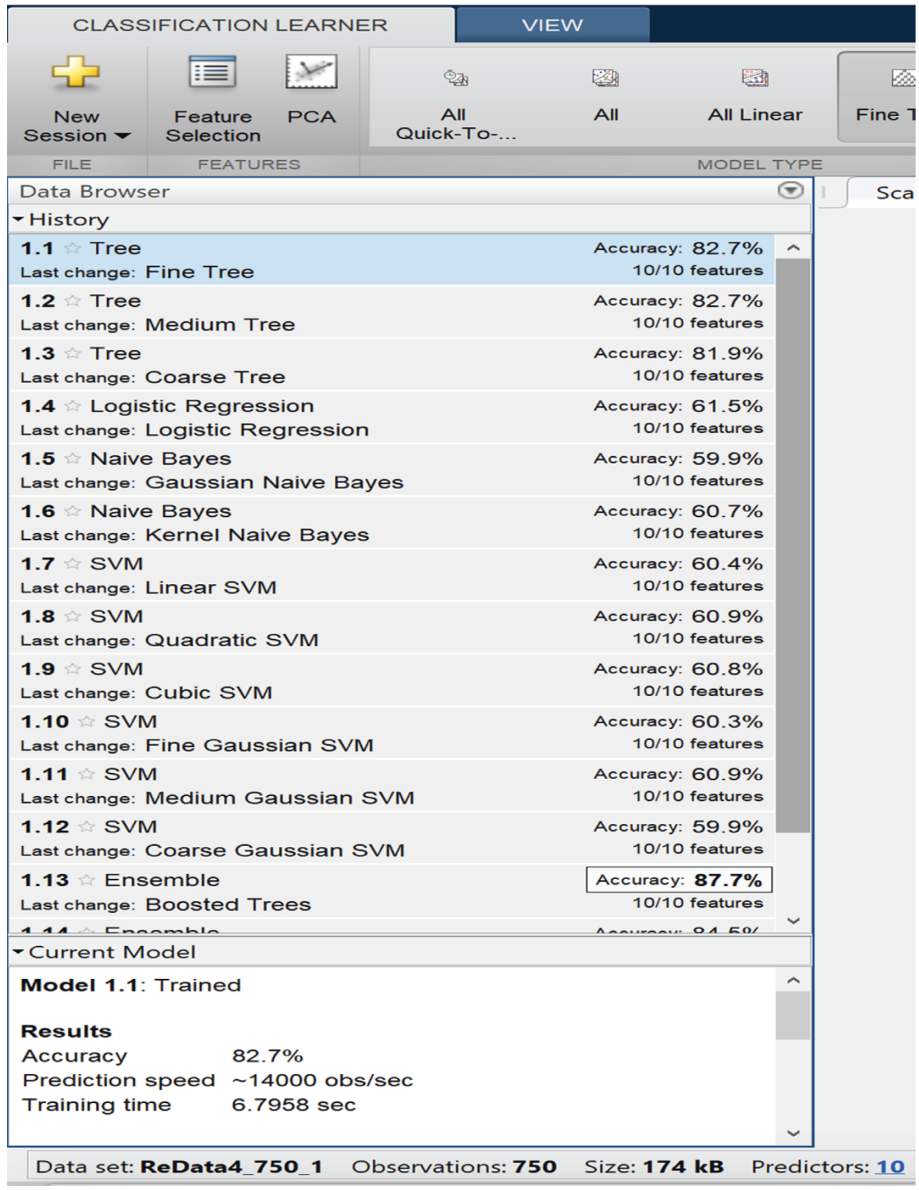

Figure 19: Algorithmic runs on 750-claim dataset.

algorithms implemented. The data used for the GSVM were subsequently used in the Classification Learner App, as shown below.

Figures 17 and 18 show the classification learner app with the various implemented algorithms and corresponding accuracies in MATLAB technical computing language environment and the results obtained using the 500-claim dataset, respectively. Figures 19 and 20 depict the subsequent results when the 750- and 1000-claim datasets were utilized for the algorithmic runs and reproducible comparison, respectively. The summarized results and accuracies are illustrated in Table 7. The summarized results in Table 7 portray the effectiveness of our proposed approach of using the genetic support vector machines (GSVMs) for fraud detection of insurance claims. From the result, it is evident that GSVM achieves a higher level of accuracy compared to decision trees and Naïve–Bayes.

5. Conclusions and Recommendations

This work aimed at developing a novel fraud detection model for insurance claims processing based on genetic support vector machines, which hybridizes and draws on the strengths of both genetic algorithms and support vector machines. The GSVM has been investigated and applied in the development of HICFDS. This paper used GSVM for detection of anomalies and classification of health insurance claims into legitimate and fraudulent claims. SVMs have been considered preferable to other classification techniques due to several advantages. They enable separation (classification) of claims into legitimate and fraudulent using the soft margin, thus accommodating updates in the generalization performance of HICFDS. With other notable advantages, it has a nonlinear dividing

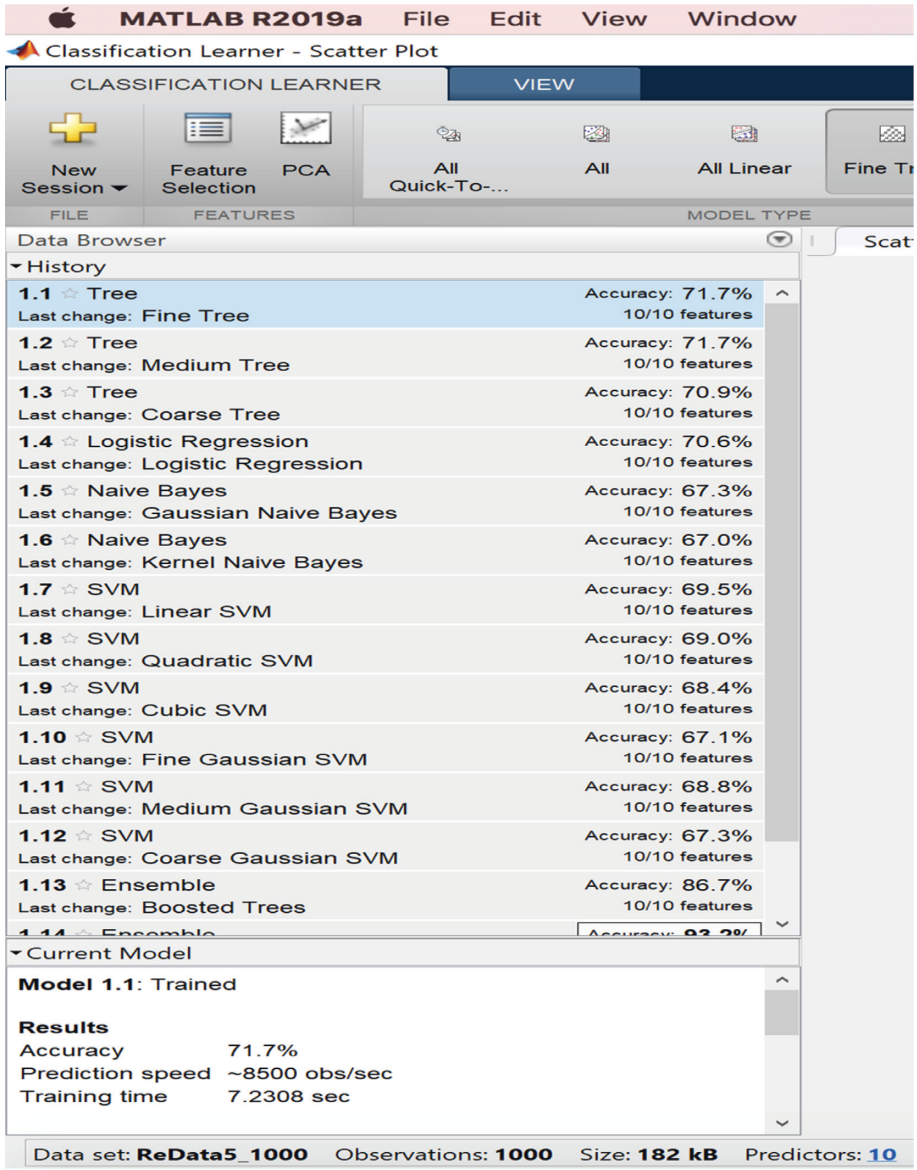

Figure 20: Algorithmic runs on the 1000-claim dataset.

hyperplane, which prevails over the discrimination within the dataset. The generalization ability of any newly arrived data for classification was considered over other classification techniques.

Thus, the fraud detection system provides a combination of two computational intelligence schemes and achieves higher fraud detection accuracy. The average classification accuracies achieved by the SVCs are 80.67%, 81.22%, and 87.91%, which show the performance capability of the SVCs model. These classification accuracies are obtained due to the careful selection of the features for training and developing the model as well as fine-tuning the SVCs' parameters using the V-fold cross-validation approach. These results are much better than those obtained using decision trees and Naïve–Bayes.

The average sample dataset testing results for the proposed SVCs vary due to the nature of the claims dataset used. This is noted in the cluster of the claims dataset (MDC specialty). When the sample dataset is much skewed to one MDC specialty (e.g., OPDC), the performance of the SVCs could tune to one classifier, especially the linear SVM, as compared to others. Hence, the behaviour of the dataset has a significant impact on classification results.

Based on this work, the developed GSVM model was tested and validated using HIC data. The study sought to obtain the best performing classifier for analyzing the health insurance claims datasets for fraud. The RBF kernel was adjudged the best with an average accuracy rate of 87.91%. The RBF kernel is therefore recommended.

Data Availability

The data used in this study are available upon request. The data can be uploaded when required.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors of this paper wish to acknowledge the Carnegie Corporation of New York through the University of Ghana, under the UG-Carnegie Next Generation of Academics in Africa project, for organizing Write Shops that led to the timely completion of this paper.

Supplementary Materials

The material consists of MS Excel file data collected from some NHIS-approved hospitals in Ghana concerning insurance claims. Its insurance claims dataset used for testing and implementation. (*Supplementary Materials*)

References

[1] G. of Ghana, National Health Insurance Act, Act 650, 2003, Ghana, 2003.

[2] Capitation, National Health Insurance Scheme, 2012, http://www.nhis.gov.gh/capitation.aspx.

[3] ICD-10 Version:2016, http://apps.who.int/classifications/icd10/browse/2016/en.

[4] T. Olson, *Examining the Transitional Impact of ICD-10 on Healthcare Fraud Detection*, College of Saint Benedict/Saint John's University, Collegeville, MN, USA, 2015.

[5] News Ghana, *NHIS Manager Arrested for Fraud | News Ghana*, News Ghana, Accra, Ghana, 2014, https://www.newsghana.com.gh/nhis-manager-arrested-for-fraud/.

[6] BioClaim Files, http://www.bioclaim.com/Fraud-Files/.

[7] Graphics Online, *Ghana news: Dr. Ametewee Defrauds NHIA of GH¢415,000—Graphic Online*, Graphics Online, Accra, Ghana, 2015, http://www.graphic.com.gh/news/general-news/dr-ametewee-defrauds-nhia-of-gh-415-000.html.

[8] W.-S. Yang and S.-Y. Hwang, "A process-mining framework for the detection of healthcare fraud and abuse," *Expert Systems with Applications*, vol. 31, no. 1, pp. 56–68, 2006.

[9] G. C. van Capelleveen, *Outlier Based Predictors for Health Insurance Fraud Detection within U.S. Medicaid*, University of Twente, Enschede, Netherlands, 2013.

[10] Y. Shan, D. W. Murray, and A. Sutinen, "Discovering inappropriate billings with local density-based outlier detection method," in *Proceedings of the Eighth Australasian Data Mining Conference*, vol. 101, pp. 93–98, Melbourne, Australia, December 2009.

[11] L. D. Weiss and M. K. Sparrow, "License to steal: how fraud bleeds America's health care system," *Journal of Public Health Policy*, vol. 22, no. 3, pp. 361–363, 2001.

[12] P. Travaille, R. M. Müller, D. Thornton, and J. Van Hillegersberg, "Electronic fraud detection in the U.S. Medicaid healthcare program: lessons learned from other industries," in *Proceedings of the 17th Americas Conference on Information Systems (AMCIS)*, pp. 1–11, Detroit, Michigan, August 2011.

[13] A. Abdallah, M. A. Maarof, and A. Zainal, "Fraud detection system: a survey," *Journal of Network and Computer Applications*, vol. 68, pp. 90–113, 2016.

[14] A. K. I. Hassan and A. Abraham, "Computational intelligence models for insurance fraud detection: a review of a decade of research," *Journal of Network and Innovative Computing*, vol. 1, pp. 341–347, 2013.

[15] E. Kirkos, C. Spathis, and Y. Manolopoulos, "Data Mining techniques for the detection of fraudulent financial statements," *Expert Systems with Applications*, vol. 32, no. 4, pp. 995–1003, 2007.

[16] H. Joudaki, A. Rashidian, B. Minaei-Bidgoli 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, 2015.

[17] V. Rawte and G. Anuradha, "Fraud detection in health insurance using data mining techniques," in *Proceedings of the 2015 International Conference on Communication, Information & Computing Technology (ICCICT)*, pp. 1–5, Mumbai, India, January 2015.

[18] J. Li, K.-Y. Huang, J. Jin, and J. Shi, "A survey on statistical methods for health care fraud detection," *Health Care Management Science*, vol. 11, no. 3, pp. 275–287, 2008.

[19] Q. Liu and M. Vasarhelyi, "Healthcare fraud detection: a survey and a clustering model incorporating geo-location information," in *Proceedings of the 29th World Continuous Auditing and Reporting Symposium*, Brisbane, Australia, November 2013.

[20] T. Ekin, F. Leva, F. Ruggeri, and R. Soyer, "Application of Bayesian methods in detection of healthcare fraud," *Chemical Engineering Transactions*, vol. 33, pp. 151–156, 2013.

[21] Home—The NHCAA, https://www.nhcaa.org/.

[22] S. Viaene, R. A. Derrig, and G. Dedene, "A case study of applying boosting naive Bayes to claim fraud diagnosis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 5, pp. 612–620, 2004.

[23] Y. Singh and A. S. Chauhan, "Neural networks in data mining," *Journal of Theoretical and Applied Information Technology*, vol. 5, no. 1, pp. 37–42, 2009.

[24] D. Tomar and S. Agarwal, "A survey on data mining approaches for healthcare," *International Journal of Bio-Science and Bio-Technology*, vol. 5, no. 5, pp. 241–266, 2013.

[25] P. Vamplew, A. Stranieri, K.-L. Ong, P. Christen, and P. J. Kennedy, "Data mining, and analytics 2011," in *Proceedings of the Ninth Australasian Data Mining Conference (AusDM'A)*, Australian Computer Society, Ballarat, Australia, December 2011.

[26] K. S. Ng, Y. Shan, D. W. Murray et al., "Detecting non-compliant consumers in spatio-temporal health data: a case study from medicare Australia," in *Proceedings of the 2010 IEEE International Conference on Data Mining Workshops*, pp. 613–622, Sydney, Australia, December 2010.

[27] J. F. Roddick, J. Li, P. Christen, and P. J. Kennedy, "Data mining and analytics 2008," in *Proceedings of the 7th Australasian Data Mining Conference (AusDM 2008)*, vol. 87, pp. 105–110, Glenelg, South Australia, November 2008.

[28] C. Watrin, R. Struffert, and R. Ullmann, "Benford's Law: an instrument for selecting tax audit targets?," *Review of Managerial Science*, vol. 2, no. 3, pp. 219–237, 2008.

[29] F. Lu and J. E. Boritz, "Detecting fraud in health insurance data: learning to model incomplete Benford's Law distributions," in *Machine Learning*, J. Gama, R. Camacho, P. B. Brazdil, A. M. Jorge, and L. Torgo, Eds., pp. 633–640, Springer, Berlin, Heidelberg, 2005.

[30] P. Ortega, C. J. Figueroa, and G. A. Ruz, "A medical claim fraud/abuse detection system based on data mining: a case study in Chile," *DMIN*, vol. 6, pp. 26–29, 2006.
[31] T. Bäck, J. M. De Graaf, J. N. Kok, and W. A. Kosters, *Theory of Genetic Algorithms*, World Scientific Publishing, River Edge, NJ, USA, 2001.
[32] M. Melanie, *An Introduction to Genetic Algorithms*, The MIT Press, Cambridge, MA, USA, 1st edition, 1998.
[33] D. Goldberg, *Genetic Algorithms in Optimization, Search and Machine Learning*, Addison-Wesley, Reading, MA, USA, 1989.
[34] J. Wroblewski, "Theoretical foundations of order-based genetic algorithms," *Fundamental Informaticae*, vol. 28, no. 3-4, pp. 423–430, 1996.
[35] J. H. Holland, *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*, MIT Press, Cambridge, MA, USA, 1st edition, 1992.
[36] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, NY, USA, 2nd edition, 2000.
[37] J. Salomon, *Support Vector Machines for Phoneme Classification*, University of Edinburgh, Edinburgh, UK, 2001.
[38] J. Platt, *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines*, Microsoft Research, Redmond, WA, USA, 1998.
[39] J. Platt, "Using analytic QP and sparseness to speed training of support vector machines," in *Proceedings of the Advances in Neural Information Processing Systems*, Cambridge, MA, USA, 1999.
[40] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, *A Practical Guide to Support Vector Classification*, Data Science Association, Taipei, Taiwan, 2003.
[41] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.
[42] D. Dvorski, *Installing, Configuring, and Developing with XAMPP*, Ski Canada Magazine, Toronto, Canada, 2007.
[43] *MATLAB Classification Learner App: MATLAB Version 2019a*, Mathworks Computer Software Company, Natick, MS, USA, 2019, http://www.mathworks.com/help/stats/classification-learner-app.html.