International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

# Catalyzing Diabetes Prediction: Harnessing Machine Learning and Deep Learning for Optimization and Clustering

**Monelli Ayyavaraiah**[1*]

**Submitted**: 05/02/2024 **Revised**: 13/03/2024 **Accepted**: 19/03/2024

**Abstract:** Diseases like diabetes mellitus are highly worrying since they kill so many people every year. High blood sugar levels are the root cause of this chronic condition. Untreated diabetes just adds extra complications to the lives of those who have it. Therefore, the mortality rate of humans may be lowered by the early prediction of diabetes. Diabetes may be better diagnosed using the data mining approach. Data mining methods for early prediction and illness detection described in a number of publications have varying degrees of accuracy. At the same time, data security is a major concern when mining information on diabetes. To address this problem, this paper develops a novel model for accurate early prediction of diabetes. In the first phase of the study, improved principal component analysis is investigated for its potential use in extracting useful features from the dataset. The machine learning approach proposes a Modified Support Vector Machine (MSVM) to diagnose diabetes at an early stage since it has the best accuracy of classification. Mining the patient's illness findings in the cloud is the key contribution of this study. The honey bee encryption and decryption algorithm is employed for this purpose. The accuracy, sensitivity, specificity, precision, and Negative Predictive Value (NPV) of the suggested approach are assessed using a number of different metrics. The collected results demonstrate the superiority of the suggested MSVM classifier, with an accuracy of 97.13%. The superior performance of the suggested approach has been shown by comparing it to the state of the art.

***Keywords:*** *Diabetes Prediction, Early Diagnosis, Data Mining, Machine Learning, Healthcare Analytics, Patient Data Security*

## Introduction

According to the 2018 World Health Organization (WHO) Report, diabetes was affecting 422 million individuals globally, positioning it as a prominent contributor to global mortality. Diabetes has a lengthy asymptomatic period; hence early diagnosis is usually desired for optimal clinical outcomes. About half of all persons with diabetes go undiagnosed because they remain in the early asymptomatic stage for too long. Only by correctly identifying typical and atypical indication signs noticed at various stages from disease onset to diagnosis can effective detection of diabetes be performed. In most cases, elevated BS levels are the root cause of the diabetic condition. Instead, it is possible that insulin production is insufficient. Recent years have seen a global increase in the number of people diagnosed with diabetes. This problem will surely be taken more seriously over the next days to guarantee that the overall number of people with diabetes decreases. Recently, several research groups have conducted in-depth analyses of one another's data mining infrastructure.

Parametric analysis of health data, such as patient sets for people with diabetes, may be utilized in data mining (DM) to synthesize knowledge in the field. In this study, we use the Deep VGG Net classifier and the adaptive gradient density-based Iterative cuckoo search optimization clustering method to analyze a set of medical data pertaining to the diagnosis of diabetes. Mining the patient's illness findings in the cloud is the key contribution of this study. Algorithms with improved ELGamal security were applied for cloud protection.

Mellitus type-2 diabetes is characterized by chronically high blood glucose (BG) levels. One of diabetes mellitus's defining features is that almost half of all diabetics have inherited the disease from a relative. Diabetes mellitus has two pathogenic causes: insufficient insulin production by the pancreas and inefficient insulin use by the organism. Diabetes mellitus may manifest in two main forms. Damaged beta cells secreted by the pancreas are the root cause of T1DM, which prevents blood glucose (BG) levels from dropping. Insulin resistance and insulin secretion failure are at the heart of Type 2 Diabetes Mellitus (T2DM), also known as non-insulin-dependent diabetes mellitus. Over the last three decades, the growing prevalence of diabetes in China has brought into focus the far-reaching effects of the disease on every aspect of Chinese society. The prevalence of diabetes continues to rise, with male cases increasing more rapidly than female ones. In 2017, approximately 110 million people in China were diagnosed with

[1*]*Assistant Professor, Department of CSE, Rajeev Gandhi Memorial College Of Engineering & Technology, Nandyal, AP, India*
***Corresponding Author:** Monelli Ayyavaraiah*
**Assistant Professor, Department of CSE, Rajeev Gandhi Memorial College Of Engineering & Technology, Nandyal, AP, India*

diabetes, according to official government statistics. This means that China has the greatest prevalence of diabetes in the world. The International Diabetes Federation (IDF) receives information on diabetes mellitus from the diabetes atlas (Seventh Edition). According to these numbers, there were 415 million people with diabetes in the globe in 2015. One in ten adults, or 642 million people, is expected to develop diabetes based on current projections of the disease's prevalence. Therefore, it is necessary to target a high-risk population of persons with diabetes mellitus in order to lessen the disease's impact and morbidity. In accordance with the most recent WHO recommendation, we define those at increased risk for developing diabetes mellitus as follows:

- Age ≥ 45 and seldom exercising
- BMI ≥ 24 kg/m2
- Impaired glucose tolerance (IGT) or impaired fasting glucose (IFG)
- Family history of diabetes mellitus
- Lower high-density lipoprotein cholesterol or hypertriglyceridemia
- (HTG) Hypertension or cardiovascular and cerebrovascular disease
- Gestation female whose age ≥30

The population at high risk for developing diabetes mellitus may be studied with the use of cutting-edge data analysis tools. The study of data mining (DM) is also timely now. Data mining, also known as database information discovery, is a computerized approach to identifying models in large data sets using a combination of artificial intelligence (AI) techniques, including machine learning, statistics, and database systems. The main objectives of these methods are to recognize patterns, draw conclusions, find correlations, and form clusters. The DM includes various processes that may be fully or partially automated to get intriguing, unknown, hidden aspects from the vast volumes of data. High-quality data and a well-executed procedure are two cornerstones of DM.

Data processing has proven useful in many areas of human existence, such as weather prediction, market analysis, medical diagnosis, and customer service. Disease monitoring and healthcare data processing will both see improvements. Patients at different hospitals have different medical histories and records. To better aid clinical research and diagnosis, it is required to assess, implement, and draw valuable knowledge from this data. It's reasonable to suppose and wait for the researchers to investigate the many promising relationships.

It's no secret that diabetes affects a huge and rising population. As a result, the vast majority of people have a superficial understanding of their own health. Therefore, it is crucial to create a model that can distinguish between presumed and confirmed high-risk diabetes mellitus patients using the diabetic data. This script uses the Deep VGG NET to predict and categorize diabetes using a flexible gradient density-based iterative Cuckoo Search Optimization (CSO) technique for clustering diabetic data. Finally, for the sake of safe conversation with the help of ELGamal encryption and decryption, with certain refinements, are used.

## I. Literature survey

Everything from serum asymmetric diabetic mellitus dimethylarginine (ADMA) to skin temperature has been evaluated. There were the everyday folks, and then there were the difficult ones. A non-contact thermo graphic camera was used to acquire thermo grams of both areas of the body. Thyroid hormones and other blood parameters are measured physiologically. Diabetic illness was included in the evaluation of the chance score. The average and highest skin temperatures in healthy people were measured at the heel and the ear. Patients with diabetes had lower than usual skin temperatures everywhere over their bodies, with the exception of the nose and tibia, where they were somewhat higher than average.

When medical data lacks a clear label that can be utilized by DM algorithms, the Semantic web has been used to express the ideas in the database. The patient information may be converted into labelled ideas that are just as good as data in a database, if not better, using the semantic web, a very advanced way of knowledge representation. Medical data is notoriously difficult to get significant insights from due to its complex and varied nature. To facilitate communication amongst hospitals' data repositories, a semantic web-based ontology-based approach has been implemented. Different types of data repositories may be involved here. The data is raised to the ontology level via the use of Resource Description Framework and Extensible Mark-up Language annotations, where it forms the domain knowledge of the hospital data store. In order to give patients with individualized healthcare, context-aware framework ontology has been created, where rules are utilized to transform low-level context into high-level context. It is possible to integrate two healthcare ontologies into a single ontology that incorporates the best features of each while eliminating redundancies.

Different kernels' potential for realizing nonlinear discriminating boundaries has been explored. The result of this data-driven preprocessing of high-dimensional patterns is a nonlinear SVM technique, the dimension of which is often substantially more than that of the original vector feature space. SVM performance is sensitive on

the selection of kernel and SVM settings. Results were evaluated using many different approaches, including re-substitution, ROC curves, learning curves, receiver operating characteristics curves, and leave-one-out-error-based 10-fold cross-validation. It is feasible to overlap stages due to the superiority of the type and form of the Kernel over the regularization constant. Because the procedure and boosting of SVMs relies on discoveries that are difficult to discern, the combination of boosts with SVMs has not improved outcomes over SVMs. This is an unconventional answer to a problem that has no obvious alternative.

An incremental Bayesian learning method that evaluates and finds improved solutions to issues is analyzed alongside the evolution of the Bayes classification via the use of a refined sample selection technique. In the sample graduation process, fresh training outcomes are analyzed for their completeness in terms of their separation, volume, and redundancy in order to inform the learning process. The emphasis is on the work put in to locate the answer above any exact calculations of how much progress has been made in each stage. The training data set may be used for incremental learning. A new approach, in comparison to traditional education, would make better use of resources like time, space, labour, and equipment by adapting a well-trained system. Problems in real-time applications may be more accurately addressed by incremental learning thanks to its focus on fine-grained analysis. It has been shown (Fan et al. 2018)

The importance of making decisions based on medical information is growing as the volume of such data continues to grow at an exponential rate. When it comes to establishing accurate classifications, decision trees are among the most dependable and productive methods available. In this work, we introduce the key elements of decision trees and explore several promising alternatives to the more familiar inductive approaches. Potential medicinal applications will be expanded as a result. The right steps are even outlined for you to follow (Cheng et al., 2020).

Rapid progress is being made in implementing CC. CC is a cutting-edge technology with endless potential for use in sectors as diverse as information technology (IT), education (E), the armed forces (AF), travel (T), culture (C), home security (I), and intelligent spaces (IS). Customers benefit greatly from the CC pricing model since they pay only for the services they really utilize. Reference: Goyal & Cumar (2019). Recent years have seen a great deal of study dedicated to the topic of cloud security. Several reliable methods have been offered by various writers to ensure the safety of data and information stored in the cloud. In this part, we'll talk about the many studies that have been conducted on the topic of cloud data center security by various researchers. Since consumers outsource their private data to cloud providers, Jouini et al. (2019) have identified data security and access control as one of the most challenging areas of current CC research. Before deciding on an encryption method, it is necessary to choose a key generation method. Several authors analyzed various strategies for key creation and administration.

## II. Proposed methodology

Table 1 shows the sample dataset's description. The Pima Indian Diabetes Dataset includes information on 768 people living in a community near Phoenix, Arizona, the United States (268 of who were diagnosed with diabetes and 500 who were not). If a person possesses a chemical that causes diabetes, it will show up as either "Tested positive" or "Tested negative." There are 8 integer properties that make up each instance. Individual medical histories and diagnostic test results are included in these files. Figure 1 is a schematic depiction of the proposed approach. The data collection has the following specific characteristics:

Number of times pregnant (preg)

Plasma glucose concentration at 2h in an oral glucose tolerance

Test (plas) Diastolic blood pressure (pres)

Triceps skinfold thickness (skin)

2-h serum insulin (insu)

Body mass index (bmi)

Diabetes pedigree function (pedi)

Age (age)

Class variable (class)

Variable type for A1 and Y is integer and from A2-A8 it is real.

**Table 1: Dataset description of Pima Indian diabetes dataset**

| Variable | Feature label | Range |
|---|---|---|
| A1 | No of pregnancy time | 0 -17 |
| A2 | The concentration of plasma glucose in 2 hr oral glucose tolerance test | 0 -199 |

| | | |
|---|---|---|
| A3 | Diastolic blood pressure | 0 -122 |
| A4 | Triceps skin fold thickness | 0 -99 |
| A5 | 2hr serum insulin | 0-846 |
| A6 | BMI | 0 - 67.1 |
| A7 | Diabetes pedigree function | 0.078 - 2.42 |
| A8 | Age | 21-81 |
| Y | Class | 0,1 |

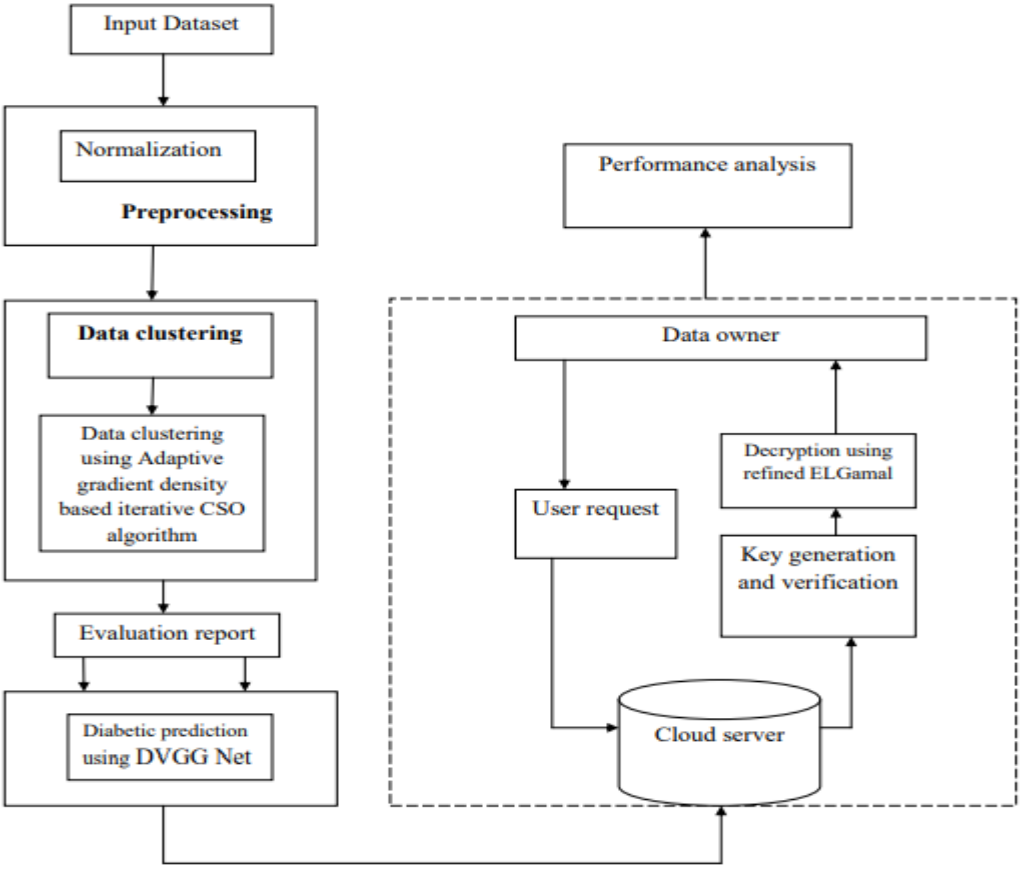


**Figure 1: Schematic representation of the suggested methodology**

**Pre-processing:** Values measured on various scales must be normalized to a nebulously standard size before the average can be calculated. In many normalizing procedures, all that's needed to get the values of the other components is a rescaling approach. When the parameters of the data are known, they may be tweaked to reduce the mistakes. In most cases, even after accounting for errors, the data values will not be distributed randomly. The z-score is the starting point for normalization.

$$Z = \left[\frac{(z-\alpha)}{\sigma}\right] \text{---------- (1)}$$

In which the data mean and standard deviation are denoted by and respectively. When both the mean and the variance are on the negative side, the final score is calculated using the sample mean and the sample variance as a whole.

$$Z = [\frac{z-z'}{\mathrm{m}}] \text{---------- (2)}$$

In which the sample mean (z') is compared to the sample standard deviation (m). In this phase of standardization, the identical values as the input values are utilized, but the mistakes are adjusted using the regression analysis formula. Consider first a basic linear regression model,

$$Y = \alpha_0 + \alpha_1 z + \epsilon \text{---------- (3)}$$

The random sample is in the form of $Y_i = \alpha_0 + \alpha_1 z_i + \epsilon_i$---------- (4)

Where $\epsilon_i$ represents the mistakes and $\sigma^2$ determines the dependence. Residuals are a kind of pseudo-error.

$$\sum_{i=1}^{n} \hat{\epsilon_i} = 0 - - - - - - - - - - - - - - - (5)$$

$$\sum_{i=1}^{n} \hat{\epsilon_i} z = 0 - - - - - - - - - - - - - - - (6)$$

Then the Hat matrix can be calculated as

$$H = X * (z^T z)^{-1} z^T - - - - - - - - - - - - (7)$$

The variance for the Hat matrix is

$$\text{Var}\left(\hat{\epsilon_i}\right) = \sigma^2(1 - h_{ii}) - - - - - - - - - (8)$$

$$\text{Var}\left(\hat{\epsilon_i}\right) = \sigma^2\left(1 - \frac{1}{n} - \left[\frac{z_i - \bar{z}^2}{\sum_{j=1}^{i}\left(z_j - \bar{z}^2\right)}\right]\right) - - - - - - - - - - (9)$$

Then the residual can be calculated by

$$t_{i=\frac{\hat{\epsilon_i}}{\hat{\sigma}}} \sqrt{1 - h_{ii}} - - - - - - - - - - - - - - - (10)$$

Where $\sigma^2$ is an estimate of the σ

$$\hat{\sigma}^2 = \frac{1}{l - m}\sum_{j=1}^{n} \bar{\epsilon}_j^2 - - - - - - - - - - - - - (11)$$

Where m is the number of parameters

**Adaptive gradient density-based iterative CSO algorithm**

When it comes time to classify the data, once it has been cleaned and prepared, we use an Iterative Cuckoo Search Optimization (ICSO) approach that is based on an adaptive gradient density. This is a computational method for investigating anomalous characteristics. A density gradient is a distribution of densities that varies over space. Any quantity whose density can be determined may be used. As a population-dependent algorithm, this one is proposed to maximize the network's usefulness. There are Cuckoo Search Optimization (CSO) guidelines available, which state that cuckoos should lay an egg in one of many possible nests. The best nest with the fewest eggs should be passed on to the following generation. The host bird will take a measurement of the cuckoo egg based on the specified number of nests and the probability distribution Pa [0, 1]. The host bird may abandon the nest or kill the cuckoo eggs if they are discovered. In order to optimize a network's parameters and its images, CSO takes a holistic approach. This function's Levy () distribution is calculated as

$$\text{Levy Class}\,(\alpha) \approx y = l - \alpha\text{---- (12)}$$

The following equation can simplify the Levy distribution

$$\theta \propto \text{Levy class}\,(\alpha) \approx \text{X} * \left(\frac{u}{|v^{1/3}|}\right)(\text{best} - x_i)\text{---------} \quad (13)$$

When calculating u and v, we use the standard deviation of these distributions and the Levy multiplication coefficient, respectively. For clustering purposes, the adaptive gradient density basis ICSO is read as follows. In the first stage, you should analyze the samples to determine their training characteristics. Cl (Cl > 2) classes are used in the approach, and it is assumed that ka is the set of pa, ma samples in the dataset's dimension. Each class between mbc and mwic has its own unique definition in the class dispatch matrix, from which the corresponding scatter matrices may be generated,

$$m_{wi_c} = \sum_{a=1}^{CL} m_a; k_a = \frac{1}{p_a}\sum_{p \in p_a} (p - Q_a)(p - Q_a)^T - - - -(14)$$

$$m_{bc} = \sum_{a=1}^{Cm} (Q_a - Q)(Q_a - Q)^T - - - - - - - - - - - - - - (15)$$

C = AT x is a dimensionality reduction formula where A is a d x d matrix. For all samples, we get the covariance, mean matrix through,

$$K = \frac{1}{n}\sum_{p \in P} (p - m)(p - m)^T - - - - - - - - (16)$$

Class distinction is enhanced by the longitudinal inequality. The significance of the covariance function and the value of its associated vector are thus determined. The score, denoted by the letter K, is calculated by taking the smaller of two vectors. By comparing the score values, we may assess the level of similarity and evaluate the query data. The deviation from ED, or Euclidean Density, is studied. The ED is shown with data that is relevant to the inquiry.

$$obj_{GDED} = -20 * \frac{q\left(-2 * \sqrt{\sum \text{k}_v}\right)}{2} - \exp\left(\frac{\sum \cos(2\pi * \text{k}_v)}{d_b}\right) + 10\text{exp} - - - - (17)$$

Where q is the data query, GDED is the Gradient Distance Euclidean Density, and k is the previous step's score value.

Algorithm 1 (Adaptive gradient density based iterative CSO)

Input: Data features $D_{n_f\ fea}$, Data_coordinate $D_c$

Output: classified valued $d_v$

To compute trust value,

For i = 1 : size $(D_{n\_fea}, 1)$

For j = 1 : size $(D_{n\_fea}, 1)$

Distance $(I, j) = \sqrt{(i_{nfea}(i,1)} - i_{nfea}(i-1)\big) + (i_{nfea}(i,1) - (i_{nfea}(i,1)^2$

End

data features $d_{n\_fea} = [d_{n\_fea}$ Distance ]

To compute, trust value $d_{tv} = (d_{n\_fea}$ dist )

Class label=unique (target)

K = length (class label)

For d = 1: k

Temp=total class mean(I, ;)

$W(I, j) = -0.5 *$ Temp* total class mean $+ \log(i)$

$W(I, 2$ :end ) = temp

Gradient density data query

$$obj_{FED} = -20 * q\left(-2 * \sqrt{\sum k_v}\right)/2 - \exp\left(\sum \cos\left(2\pi * k_v\right)/d_b\right) + 10\exp$$

End

Disease data categorization is the last stage of the proposed system. If the data has already been clustered, then classifying it will be a breeze. Using this Deep VGG Net CNN, the gap between a single metric and several others may be calculated. CNN evaluates probability and usefulness. That is the whole total of the data gathered. Here, CNN will act as an interpreter and redistributors of the data, before using its class probability to compute the classification.

The process is started by the activation of the neuron,

$C_j^l = \sigma_k \sum_K D_{jk}^l s_K^l + e_j^l$----- (18)

The equation can be written in the form of vector,

$$N^l = \sigma_k(d^l s^{l-1} + e^l) - - - -(19)$$

The quadratic set in which the training set can be merged,

$$C = \frac{1}{2}\|Y - s^l\|^2 = \frac{1}{2}\sum_j \left(y_j - S_j^l\right)\text{^}2 - (20)$$

The gradient output is given by,

$$\frac{\partial s}{s} = ek^{1-1}\delta_j^l = C$$

$$C = \det[\text{ features }] - k(\text{ classify }(c))^2 - (21)$$

The characteristics that have been categorised are denoted by C, and the pointed feature k. These must be reported as

$$\det[C] = Ne_j^l - - - (22)$$

$$\text{classify}(c) = ke_j^l - - - (23)$$

The CNN classification was concluded as

$C = Ne_j^1 - m\left(N_j^1 e_j^1\right)^2$---- (24)

Algorithm 2 (Deep VGG Net CNN Classification)

Input: Processed data $P_{data}$

Output: classified data $\psi_d$

Initialize the Network layers

Initialize train features

initialize label data

Train label data = 70%

Testt label data = 30%

Label=unique (label)

For ii=1: lengh (Lab)

Class = find ( label == Lab (ii))

Train cut=length (class)-train cut data

Train data= [train data; train features; class (1: Train cut) end-6: end]

Predict labelled data=classify (net, train data) End

End For ii=4: size (traindata, 5)

Train data= [train data; train features; class (1: Train cut) end-6: end]

End

For ii=4: size (train features, 5)

Train data= [train features; train features; class (1: Train cut) end-6: end]

End

**III.Refined ELGamal algorithm**

The proposed security mechanism ensures both the security and efficiency of cloud-based data exchange. Instead of RSA, the ELGamal Algorithm offers an alternative for public-key encryption. Many contemporary Attribute-Based Encryption (ABE) strategies, with a single authority, can manage both private and public keys. However, in specific cases, system administrators need to share data with consumers who are overseen by a separate agency. To address this challenge, several multi-authority systems have been established. Data holders have access control systems to update ciphertext and attributes with similar status. The

suggested scheme incorporates a security algorithm for weighing attributes in cloud storage records.

The Refined ELGamal symmetric 64-bit Block cipher boasts a variable-length address ranging from 32 to 448 bits (equivalent to 14 bytes). This algorithm was purposefully designed to precisely and consistently encrypt 64-bit plaintext into 64-bit ciphertext. When selecting operations for the algorithm, options include table scanning, node operations, addition, and bit-by-bit processing, all aimed at minimizing the required 32-bit processor encryption and decryption time. Despite its complexity, the algorithm was intentionally crafted to maintain straightforward and simple code functions without compromising security.

Similar to the Data Encryption Standard (DES), Refined ELGamal utilizes a 16-round Feistel network for both encryption and decryption. However, unlike DES, where only the right 32 bits are modified in each round, Refined ELGamal modifies both the left and right 32-bit sections after each round. Moreover, before the modified F-function or 32-bit operation on the right in the next round, Refined ELGamal includes a bitwise exclusive OR operation on the left. Additionally, Refined ELGamal consists of two proprietary activities and an exchange process after 16 iterations. The mechanism's structure is rooted in the DES permutation format.

The suggested system design incorporates the Refined ELGamal algorithms for encoding, decryption, and key generation. In turn, the findings are verified using a code-matching procedure. Instead, the program's functioning determines how much user weight it generates. Key extension and key encryption are often treated as separate steps in the Refined ELGamal method. There are 16 circles used to encrypt the data. Every time a key or piece of data is used, it is displaced and replaced with another. This supplement is built to compete with 32 bits (four indexed search tables). Refined ELGamal displays all of these details. Certificate Authorities (CAs) are responsible for issuing a single user ID to each client that establishes network connectivity.

But the user encrypts the attributes and then signs them over to the authority. The legitimate feature verifies the user's claimed identity. Keys and authority will be transferred to the new user if they are the correct person in charge. CA and authority broadcast a secret key to the network and provide the active user with their own private secret key. With the use of the central authority's setup and software algorithms, the challenger would be given access to the correct keys, allowing them to launch an assault. Until a data file is uploaded to the server, the data manager signs in with a single ID and chooses a symmetrical dataset key at random. In the first step, the user of the computer implements cloud computing and employs a decryption technique. If the data owner's secret key is valid, the procedure uses it to determine the relative importance of each piece of information. The receiver may read the encrypted message attached to the weighted text.

## IV. Result and Discussion

In this section, we show how the suggested technique may help. As a measure of the quality of the pre-trained model, the suggested model delivers beneficial results on a set of test data. Accuracy (A), Precision (P), Recall (R), Mathew's correlation coefficient (MCC), and kappa statistic to evaluate classification output (for sample analysis) are the metrics used to determine the impacts of experimental classification. The time and effort required to encrypt and decode data is also determined. The suggested strategy outperforms previous methods while requiring less effort and may be simply implemented at scale.

### Accuracy

It is a test of mathematical propensity. Accuracy may be defined as the fraction of total data that consists of real findings (both positive and negative).

$$Accuracy = (TP + TN)/(TP + TN + FP + FN)$$

### Precision

The precision can be calculated by using the equation,

$$\text{precision } = \frac{TP}{TP + FP}$$

Recall The recall can be calculated by using the equation (4.29),

$$\text{Recall } = TP/(TP + FN)$$

### MCC

For a quantitative evaluation of the precision of your classifications, use the formula to get the Mathews correlation coefficient.

$$\frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}}$$

### Kappa statistic

When evaluating the precision of a model, the kappa statistic is a crucial indicator. The results of the proposed model are compared to those of the randomly selected procedure. Values for kappa varied from zero to one. Values close to 1 represent the expected model impact, whereas values close to 0 represent no effect.

$$\mathrm{K} = [\mathrm{p(a)} - \mathrm{p(e)}]/[1 - \mathrm{p(e)}]$$

$$p(a) = (TP + TN)/N$$

$$p(e) = [(TP + FN) * (TP + FP) * (TN + FN)]N\text{^}2$$

The rating results are determined by these markers of performance. Important performance is prioritized via the solutions' connections to related algorithms and categorization methods. Due to the superior performance of VGG NET CNN, the suggested classifier can get the best results compared to state-of-the-art methods for diagnosing diabetes and making long-term predictions. There are a total of 300 cases in the dataset, split evenly between positive and negative examples. We gained insight into the utility of the suggested approach by testing it on a real-world dataset. Experiment results confusion matrix is shown in Table 2 diabetes categories in one column (A) and the severity levels in the first row (B to E)

**Table 2: Confusion matrix**

| A | B | C | D | E |
|---|---|---|---|---|
| Early diabetes | 114 | 0 | 0 | 0 |
| Intermediate diabetes | 0 | 23 | 0 | 0 |
| Late diabetes | 0 | 0 | 100 | 0 |
| Severe diabetes | 0 | 0 | 8 | 55 |

The dataset provided several noteworthy findings after going through preprocessing and classification, as seen in Table 3 and Figure 2. The suggested model was put to the test on the dataset by contrasting it with the standard approach [Han Wu et al., 2016]. The outcome is shown in the following example. Predictive accuracy was around 98.6%, demonstrating the validity and utility of the suggested model.

**Table 3: Comparative analysis**

| Performance metrics | Improved K-means with logistic regression (Existing ) | Proposed |
|---|---|---|
| Accuracy | 93% | 98.6% |
| Precision | 92.5% | 99.7% |
| Recall | 92.9% | 95% |
| MCC | 78.6% | 89% |
| ROC area | 96.2% | 97.3% |
| Kappa statistic | 78.6% | 89% |

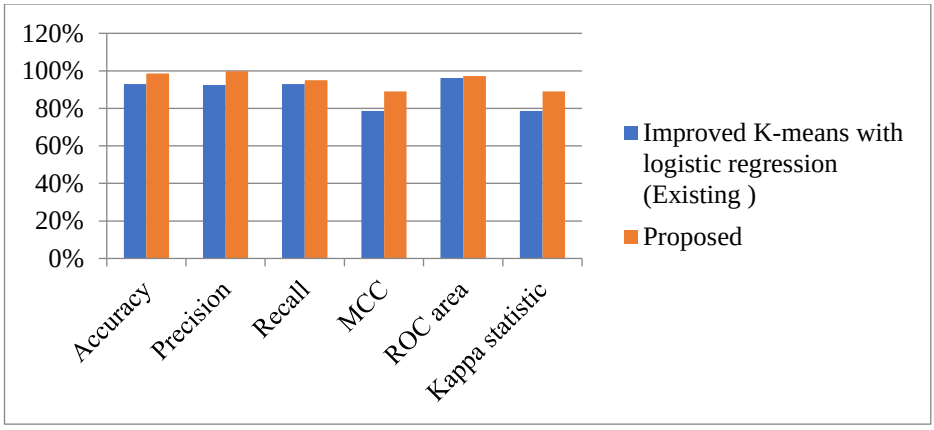


**Figure 2: Comparative analysis**

This representation illustrates the encryption time required for original medical data. As the volume of medical data grows, the encryption time also experiences a corresponding increase. The performance metrics of the proposed cloud security solution, the Refined ELGamal algorithm, were assessed based on the time taken by this algorithm for both encryption and decryption. These time measurements are detailed in Tables 4 and 5, allowing for a comparison with contemporary security algorithms to substantiate the effectiveness of the Refined ELGamal algorithm.

**Table 4 Comparison of encryption time**

| Algorithm | 20KB | 50KB | 150KB | 300KB |
|---|---|---|---|---|
| MDRSA (Kiran et al.2020) | 0.051 | 0.078 | 0.093 | 0.25 |
| DES (Nora et al. 2019) | 0.038 | 0.042 | 0.08 | 0.129 |
| AES (Nora et al. 2019) | 0.027 | 0.037 | 0.06 | 0.076 |
| ELGamal | 0.012 | 0.016 | 0.018 | 0.065 |

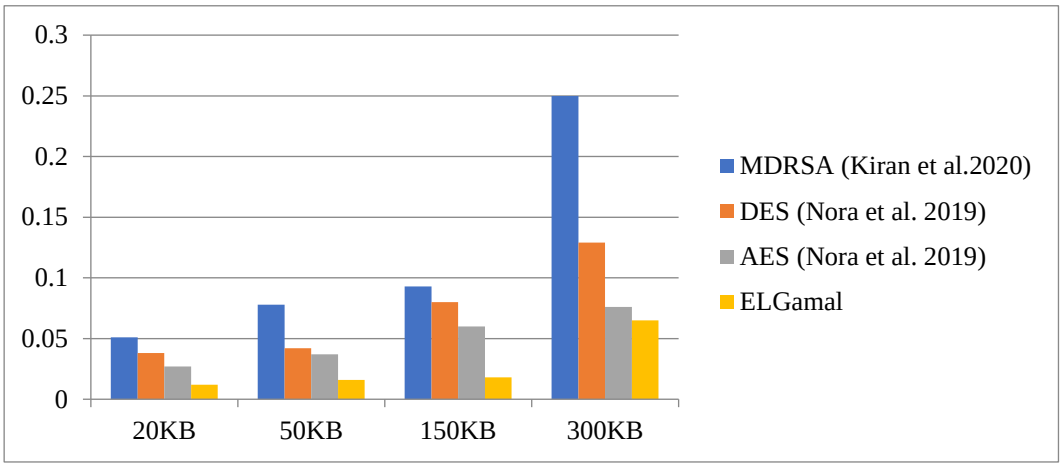


**Figure 3: Time taken for encryption**

The re-encrypted ciphertext undergoes decryption using the private key of the healthcare professional. Table 5 provides insights into the decryption time concerning the re-encrypted data in relation to the re-encrypted file size. Meanwhile, Figure 4 portrays the simulation specifics regarding the decryption time.

**Table 4 Comparison of Decryption time**

| Algorithm | 20KB | 50KB | 150KB | 300KB |
|---|---|---|---|---|
| MDRSA (Kiran et al.2020) | 0.035 | 0.05 | 0.085 | 0.1 |
| DES (Nora et al. 2019) | 0.005 | 0.012 | 0.036 | 0.076 |
| AES (Nora et al. 2019) | 0.011 | 0.017 | 0.028 | 0.048 |
| ELGamal | 0.009 | 0.0092 | 0.013 | 0.02 |

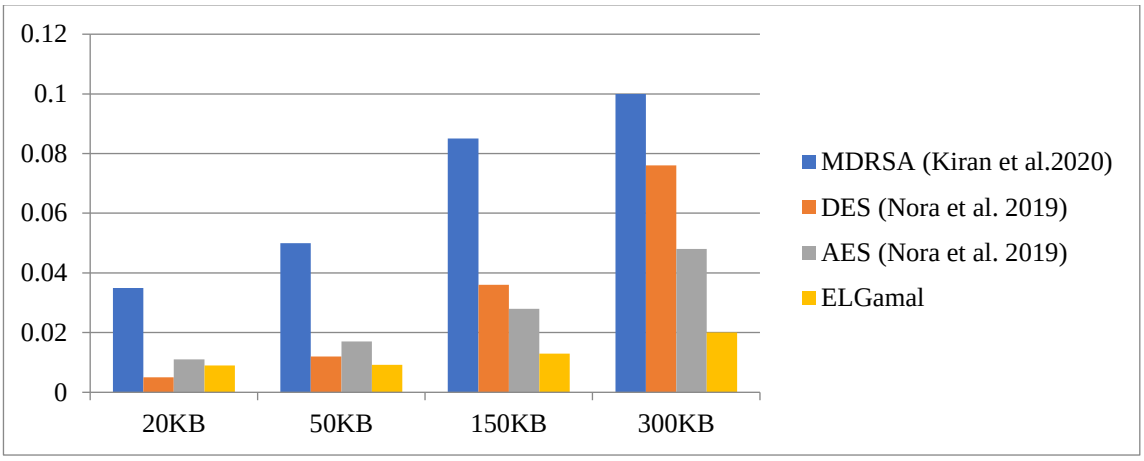


**Figure 4 Time taken for decryption**

Figures 3 and 4 demonstrate that when encrypting and decrypting data sizes ranging from 20KB to 300KB, Refined ELGamal outperforms AES, DES, and MDRSA. After subjecting all four algorithms to a rigorous evaluation with files of various sizes, we observed that our proposed technique encrypts and decrypts data at a faster pace compared to the competition. In contrast to other approaches, the suggested method exhibits a significant speed improvement as file sizes increase. This highlights the superiority of the improved ELGamal as a faster, more secure, and visually appealing cryptographic method.

## V. Conclusion

Among these experiments, the most significant challenge in the medical sector lies in the assessment and treatment of diabetes using current methods. To address this challenge, we employed Deep VGG Net for the classification of diabetes data. The primary objective of this investigation is to facilitate early and effective diabetes detection, thus saving both time and resources.

The proposed model encompasses both data prediction and cloud data storage, with a focus on the analysis of the publicly available PIMA Diabetes dataset. Remarkably, our model achieved an impressive predictive accuracy of 98.6%. Additionally, in evaluating cloud data security, we compared our proposed Refined ELGamal algorithm with the Honey Pot security framework. The results unequivocally demonstrate that Refined ELGamal effectively ensures high data security.

This patient information can be stored in a database for further data analysis and the development of enhanced layout improvement techniques. Such an approach not only contributes to individual well-being but also helps establish healthier living conditions overall.

## References

[1] Aljawarneh, S & Yassein, MB 2017, 'A resource-efficient encryption algorithm for multimedia big data', Multimedia Tools and Applications, vol. 76, no. 21, pp. 22703-22724. [2]. Al-Sakran, HO 2015, 'Development of business analytics curricula to close skills gap for job demand in big data', Development, vol. 5, no. 3.

[2] Al-Shaikhly, MH, El-Bakry, HM & Saleh, AA 2018, 'Cloud security using Markov chain and genetic algorithm', International Journal of Electronics and Information Engineering, vol. 8, no. 2, pp. 96-106.

[3] Amani Yahyaoui, Akhtar Jamil, Jawad Rasheed & Mirsat Yesiltepel 2019, 'A Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Techniques', 1st International Informatics and Software Engineering Conference.

[4] Ashari, A, Paryudi, I & Tjoa, AM 2013, 'Performance comparison between Naïve Bayes, decision tree and k-nearest neighbor in searching alternative design in an energy simulation tool', International Journal of Advanced Computer Science and Applications (IJACSA), vol. 4, no. 11.

[5] Behbahani, BA, Yazdi, FT, Shahidi, F, Mortazavi, SA & Mohebbi, M 2017, 'Principle component analysis (PCA) for investigation of relationship between population dynamics of microbial pathogenesis, chemical and sensory characteristics in beef slices containing Tarragon essential oil', Microbial pathogenesis, vol. 105, pp. 37-50.

[6] Belguith, S, Kaaniche, N, Laurent, M, Jemai, A & Attia, R 2018, 'Phoabe: Securely outsourcing multi-authority attribute based encryption with policy hidden for cloud assisted iot', Computer Networks, vol. 133, pp. 141-156.

[7] Beyene, C & Kamat, P 2018, 'Survey on prediction and analysis the occurrence of heart disease using data mining techniques', International Journal of Pure and Applied Mathematics, vol. 118, no. 8, pp. 165-174.
[8] Cheng, X, Chen, F, Xie, D, Sun, H & Huang, C 2020, 'Design of a secure medical data sharing scheme based on blockchain', Journal of medical systems, vol. 44, no. 2, pp. 1-11.
[9] Choi, C, Choi, J & Kim, P 2014, 'Ontology-based access control model for security policy reasoning in cloud computing', The Journal of Supercomputing, vol. 67, no. 3, pp. 711-722.
[10] Cios, KJ & Moore, GW 2002, 'Uniqueness of medical data mining', Artificial intelligence in medicine, vol. 26, no. 1-2, pp. 1-24.
[11] Deepa, N & Pandiaraja, P 2020, 'E health care data privacy preserving efficient file retrieval from the cloud service provider using attribute based file encryption', Journal of Ambient Intelligence and Humanized Computing, pp. 1-11.
[12] Devi, MR 2016, 'Analysis of various data mining techniques to predict diabetes mellitus'.
[13] Devi, TD, Subramani, A & Anitha, P 2020, 'Modified adaptive neuro fuzzy inference system based load balancing for virtual machine with security in cloud computing environment', Journal of Ambient Intelligence and Humanized Computing, pp. 1-8.
[14] Dey, M & Rautaray, SS 2014, 'Study and analysis of data mining algorithms for healthcare decision support system', planning, vol. 5, no. 6.
[15] Durairaj, M & Kalaiselvi, G 2015, 'Prediction of diabetes using soft computing techniques-A survey', International journal of scientific & technology research, vol. 4, no. 3, pp. 190-192.
[16] Ephzibah, E 2011, 'Cost effective approach on feature selection using genetic algorithms and fuzzy logic for diabetes diagnosis', arXiv preprint arXiv:1103.0087.
[17] Evirgen, H & Çerkezi, M 2014, 'Prediction and Diagnosis of Diabetic Retinopathy using Data Mining', The Online Journal of Science and Technology, vol. 12, p. 32.
[18] Fan, K, Wang, S, Ren, Y, Li, H & Yang, Y 2018, 'Medblock: Efficient and secure medical data sharing via blockchain', Journal of medical systems, vol. 42, no. 8, pp. 1-11.
[19] Fan, R-E, Chen, P-H, Lin, C-J & Joachims, T 2005, 'Working set selection using second order information for training support vector machines', Journal of machine learning research, vol. 6, no. 12.
[20] Fayyad, U, Piatetsky-Shapiro, G & Smyth, P 1996, 'From data mining to knowledge discovery in databases', AI magazine, vol. 17, no. 3, pp. 37-54.
[21] Ferrag, MA, Maglaras, L, Moschoyiannis, S & Janicke, H 2020, 'Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study', Journal of Information Security and Applications, vol. 50, p. 102419.
[22] Galathiya, A, Ganatra, A & Bhensdadia, C 2012, 'Improved decision tree induction algorithm with feature selection, cross validation, model complexity and reduced error pruning', International Journal of Computer Science and Information Technologies, vol. 3, no. 2, pp. 3427-3431.
[23] Ghosh, P, Biswas, S, Shakti, S & Phadikar, S 2020, 'An improved intrusion detection system to preserve security in cloud environment', International Journal of Information Security and Privacy (IJISP), vol. 14, no. 1, pp. 67-80.
[24] Gupta, R, Kanungo, P & Dagdee, N 2020, 'HD-MAABE: Hierarchical Distributed Multi-Authority Attribute Based Encryption for Enabling Open Access', in International Conference on Intelligent Computing and Smart Communication 2019: Proceedings of ICSC 2019, p. 183.
[25] Hepsiba, CL & Sathiaseelan, J 2016, 'Security issues in service models of cloud computing', International Journal of computer science and Mobile Computing, vol. 5, no. 3, pp. 610-615.
[26] Hossin, M & Sulaiman, M 2015, 'A review on evaluation metrics for data classification evaluations', International Journal of Data Mining & Knowledge Management Process, vol. 5, no. 2, p. 1.
[27] Hourali, M & Montazer, GA 2011, 'An intelligent information retrieval approach based on two degrees of uncertainty fuzzy ontology', Advances in Fuzzy Systems, vol. 2011.
[28] Huang, F, Huang, J & Shi, Y-Q 2016, 'New framework for reversible data hiding in encrypted domain', IEEE transactions on information forensics and security, vol. 11, no. 12, pp. 2777-2789.
[29] Iyer, A, Jeyalatha, S & Sumbaly, R 2015, 'Diagnosis of diabetes using classification mining techniques', arXiv preprint arXiv:1502.03774.