DISCOVERING THE TYPE 2 DIABETES IN EHR USING THE SB-SVM

Ms.M.ANITHA[1],Ms. A. HARIKA VENKATA SIVA NAGA SAI DURGA[2]

#1 Assistant professor in the Master of Computer Applications in the SRK Institute of Technology, Enikepadu, Vijayawada, NTR District

#2 MCA student in the Master of Computer Applications at SRK Institute of Technology, Enikepadu, Vijayawada, NTR District

ABSTRACT_ The finding of Type 2 diabetes (T2D) at a beginning phase plays a vital part for a sufficient T2D coordinated administration framework and patient's development. The World Health Organization (WHO) reported that the global prevalence of worldwide diabetes is around 9% (more than 400 million people). Recent years have witnessed an increasing amount of available Electronic Health Record (EHR) data for diabetes and Machine Learning (ML) techniques have been considerably evolving.

In particular, among all the EHR features related to exemptions, examination and drug prescriptions we have selected only those collected before T2D diagnosis from a uniform age group of subjects. Overfitting, model interpretability, and computational cost are just a few of the issues that may arise when modeling and managing this amount of data.

As a result, we developed a machine learning technique known as the Sparse Balanced Support Vector Machine (SB-SVM), which achieves the ideal balance between computation time and predictive performance. In addition, the model's interpretability is improved by the induced sparsity, which implicitly manages the high dimensional data and typical unbalanced class distribution.

1. INTRODUCTION

Diabetes is a chronic disease that occurs when the pancreas doesn’t produce insulin or when the body cannot effectively use the insulin it produces. Insulin is a hormone that regulates blood sugar. Hyperglycaemia, or raised blood sugar, is a common effect of uncontrolled diabetes and overtime leads to severe damage to many of the body systems, especially the nerves and the blood cells.

In 2010, it was estimated that 285 million people worldwide (6.4% of adults) had diabetes. That figure is projected to increase to 552 million by 2030. According to the disease's current growth rate, one in ten adults should have diabetes by the year 2040. Additionally, there has been a sharp rise in the prevalence of diabetes in South Korea, where 13.7% of adults there have the disease and nearly a quarter have prediabetes, according to recent studies.

Diabetes frequently goes undiagnosed because those who have it are frequently unaware of their condition or are themselves asymptomatic; nearly a third of diabetic patients are unaware of their condition. The kidneys, heart, nerves, blood vessels, and eyes are just a few of the body systems and organs that suffer severe, long-lasting damage from uncontrolled diabetes. Therefore, early disease detection enables those who are at risk to take preventive action to slow the disease's progression and enhance quality of life. Diabetes can be managed and complications can be avoided with regular blood sugar monitoring and early intervention through dietary changes, medication, or insulin therapy. To create a specialised treatment plan, people with diabetes must collaborate closely with their medical team.

Diabetes is regarded as a chronic condition that affects people of all ages. The exact origin of the illness is still a mystery. Age, family history, other relative illnesses, pregnancy, fluctuating glucose levels, blood pressure, etc. are some of the factors or causes, though. Diabetes is a condition that can be managed with medication, but as of now, there is no complete medical cure for the condition. Type 1 diabetes, Type 2 diabetes, gestational diabetes, and prediabetes are the four main types of diabetes. Under these four categories, there are also some sub-types. Insulin-dependent diabetes, also referred to as "Type 1 diabetes," is a condition where the insulin release cell is damaged and unable to produce insulin.

In "Type 2" diabetes, the body fails to produce enough insulin. This typically occurs at an average age of over 40 years. Pregnancy is when "gestational diabetes (GDM)" most frequently occurs. Prediabetes, the final of the four main categories, is when blood sugar levels are higher than normal but not as high as Type 2 diabetes. Prediabetes is a warning sign that indicates an increased risk of developing Type 2 diabetes and heart disease. Lifestyle changes such as losing weight, increasing physical activity, and eating a healthy diet can prevent or delay the onset of Type 2 diabetes in people with prediabetes.

2. LITERATURE SURVEY

2.1 Choi B.G., Rha S.-W., Kim S.W., Kang J.H., Park J.Y., Noh Y.-K. Machine learning for the prediction of new-onset diabetes mellitus during 5-year follow-up in non-diabetic patients with cardiovascular risks.*Yonsei Med. J.* 2019; 60:191–199.

Purpose: Predictive models for Type 2 diabetes mellitus (T2DM) have been proposed in numerous studies. However, these predictive models are constrained by a number of factors, including reproducibility and ease of use for users. Using machine learning and electronic medical records (EMRs), this study created a T2DM predictive model and compared its efficacy to that of more conventional statistical methods.

Methods and Materials: A total of 8454 patients treated at Korea University Guro Hospital's cardiovascular centre without a diabetes history were included in this study. All subjects were followed for five years. The commonness of

T2DM during follow up was 4.78% (404/8454). From the EMRs, a total of 28 variables were extracted. The logistic regression (LR), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), and K-nearest neighbour (KNN) algorithm models were created in order to confirm the prediction model-based cross validation test. The LR model was considered to be the current method of statistical analysis.

Results: After a 10-fold cross-validation test, every predictive model maintained a change within the standard deviation of the area under the curve (AUC) of 0.01. With an AUC of 0.78, the LR learning model outperformed all other predictive models in terms of prediction accuracy. Nonetheless, contrasted with the LR model, the LDA, QDA, and KNN models didn't show a genuinely massive distinction.

Conclusion: Using machine learning and an EMR database, we developed and verified a T2DM prediction system that predicted the occurrence of T2DM over the course of five years in a manner that was comparable to that of a conventional prediction model. Clinical research is required to apply and validate the prediction model in subsequent research.

2.2 Wei S., Zhao X., Miao C. A comprehensive exploration to the machine learning techniques for diabetes identification; Proceedings of the 2018 IEEE 4th World Forum on Internet of Things (WF-IoT); Singapore. 5–8 February 2018; pp. 291–295.

It presents a thorough analysis of machine learning methods for diabetes detection. For a variety of machine learning algorithms, the study examined two crucial data processors: PCA (Principal Component Analysis) and LDA (Linear Discriminant Analysis). They carried out parameter tuning to find the best performance after conducting an experiment to determine the best data pre-processor for each algorithm. The Pima Indian data set was used to test the algorithms' effectiveness. Using 10-fold cross-validation, the accuracy that was obtained using the five algorithms (neural network, support vector machine, decision tree, logistic regression, and naive bayes) that were used was 77.86%. According to this study, using more sophisticated techniques, such as ensemble methods or deep learning models, could increase the algorithms' accuracy. Additional study might look into the use of various data sets to evaluate how well these algorithms perform across various domains. Additionally, it's crucial to think about the ethical ramifications of using these algorithms because, if not properly developed and tested, they have the potential to reinforce prejudices and discrimination. Future studies should concentrate on creating transparent and fair algorithms that can be applied in various contexts.

Diabetes mellitus, known as diabetes, is a gathering of metabolic problems and has impacted many millions of individuals. The location of diabetes is vital, concerning its serious confusions. Numerous studies on diabetes identification have been conducted, many of which are based on the Pima Indian diabetes data set. Beginning in 1965, this data set examines women in the Pima Indian population, where the onset rate of diabetes is relatively high. The majority of previous research focused primarily on a few distinct complex techniques for testing the data; however, there has not been a comprehensive examination of many common techniques.

3. PROPOSED SYSTEM

To implement this model, we would Data Gathering Model building, pre-processing, prediction, and evaluation. To achieve the highest level of accuracy, the proposed system combines the benefits of feature selection and SB-SVM to identify Type 2 Diabetes in EHR. The process of feature selection contributes to the reduction of noise and the number of unimportant features in the data, which enhances the performance of the SB-SVM model. This system has the potential to help medical staff correctly identify patients with Type 2 Diabetes, resulting in better disease management and treatment. Additionally, the SB-SVM model can be trained on a sizable dataset, enabling more precise predictions and individualised treatment plans for Type 2 Diabetes patients. In the end, this might result in lower medical expenses and better health outcomes.

3.1 IMPLEMENTATION

Data pre-processing

Pre-processing describes the changes made to our data before we feed it to the algorithm. A technique for turning unclean data into clean data sets is data pre-processing. In other words, data is always gathered from various sources in a raw state that precludes analysis. Data pre-processing entails a number of steps, such as data cleaning, data transformation, and data reduction, to ensure that the data is accurate, consistent, and suitable for analysis. It improves the data's quality and increases the accuracy of the inferences made from it, making it a crucial step in the analysis of data.

A lot of real-world data has noises, missing values, and may be in an unusable format so that machine learning models can't directly use it. Cleaning the data and making it suitable for a machine learning model which also improves the accuracy and efficiency of a machine learning model—requires data pre-processing.

It includes beneath steps

Finding missing data,

Encoding categorical data,

Splitting the dataset into a training and test set,

Feature scaling are all steps in the import process of the dataset.

Data Visualisation

Data visualisation is the process of transforming sizable data sets into statistical and graphical representations. In data science and knowledge discovery techniques, it is essential to make data more understandable and accessible. Visual representation is required for charts and graphs in order to facilitate quick information absorption and make them easier to understand. Avoid hesitating on tables with large data sets if you want to keep the audience's interest for a longer period of time. Additionally, by using visualisation, it is possible to identify patterns and trends that may not be readily apparent when examining raw data, which can aid in problem-solving and better-informed decision-making. Additionally, it can help in simplifying complex information so that a larger audience can understand it.

Model Building

The following are the six steps to building a machine learning model:

1. Contextualize machine learning in your organisation .
2. Explore the data and choose the type of algorithm
3. Prepare and clean the dataset
4. Split the prepared dataset and perform cross validation
5. Perform machine learning optimisation
6. Deploy the model

Contextualise Machine Learning in your association

Understanding why your company requires a machine learning model is the first step in creating one. Since machine learning development can take a lot of resources, it's important to agree on and set clear goals early on. Define the problem that a model must solve and the characteristics of success in detail. A conveyed model will bring significantly more worth on the off chance that it's completely lined up with the targets of your association. There are important aspects that must be investigated and planned before the project can begin.

At this stage the accompanying subtleties ought to be concurred:

The general proprietors of ML project.

A definition of project success and the problem that the project must solve.

The sort of issue the model should tackle.

The objectives of the model to comprehend profit from speculation once sent.

The quantity and quality of training data coming from the source.

Whether it is possible to use pre-trained models instead.

The building of a machine learning model will be streamlined if a pre-trained model can be realigned and used to solve the problem. Transfer learning can use an existing model to solve a similar problem rather than building a new model from scratch. This will reduce the number of resources needed for the project, especially for supervised learning, which needs a lot of labelled training data in large arrays.

Explore the data and choose the type of algorithm:

Choosing the right kind of model is the next step in building a machine learning model. The distinctions rely upon the sort of assignment the model necessities to perform and the elements of the dataset within reach. Through exploratory data analysis, a data scientist should first investigate the data.

There are three primary types of machine learning models. Ever one has a particular way to deal with preparing the model. Labelled datasets that have been prepared by a data scientist are needed for supervised machine learning models. As a result, input and labelled output data will be included in the training dataset. Predicting outcomes and categorizing new data are two applications for supervised machine learning models. Unlabelled datasets are used to train unsupervised machine learning models.

Prepare and clean the dataset

ML models for the most part need huge varieties of great preparation information to guarantee an exact model. For the most part, the model will gain the connections among info and result information from this preparing dataset. The nature of the machine learning training that is being carried out will have an impact on the composition of these datasets. Labelled datasets, which contain both labelled input variables and labelled output variables, are used to train supervised machine learning models.

A data scientist usually completes the labour-intensive process of preparing and labelling the data. In contrast, unsupervised machine learning models will not require labelled data for their training dataset, which will only consist of input variables or features. In the two sorts of ML the nature of information significantly affects the general adequacy of the model. Since the model learns from the data, bad data may make the model ineffective when it is put into use. In order to standardize the data, find any missing data, and find any outliers, the data should be checked and cleaned.

Split the prepared dataset and perform the cross-validation

The cycle is called cross approval in ML, as it approves the viability of the model against concealed information. There are a variety of cross validation methods, which can be categorized as exhaustive or non-exhaustive methods. All possible combinations and iterations of training and testing dataset will be tested using exhaustive cross validation methods. Non-comprehensive cross approval procedures will make a randomized parcel of preparing and testing subsets. The comprehensive methodology will give more top to bottom understanding into the dataset, however will take significantly more time and assets rather than a non-thorough methodology.

Perform Machine Learning optimisation

Model hyperparameters, which are model configurations set by the data scientist, are evaluated and rearranged as part of the process of machine learning optimization. Hyperparameters aren't learned or created by the model through AI. Instead, the model's designer chose and established these configurations. The model's structure, the learning rate, or the number of clusters a model should classify data into are all examples of hyperparameters. After optimizing the hyperparameters, the model will be able to carry out its tasks with greater efficiency.

Deploy the model

The last step in building a machine learning model is the deployment of the model. Machine learning models are generally developed and tested in a local or offline environment using training and testing datasets. Deployment is when the model is moved into a live environment, dealing with new and unseen data. This is the point that the model starts to bring a return on investment to the organization, as it is performing the task it was trained to do with live data.

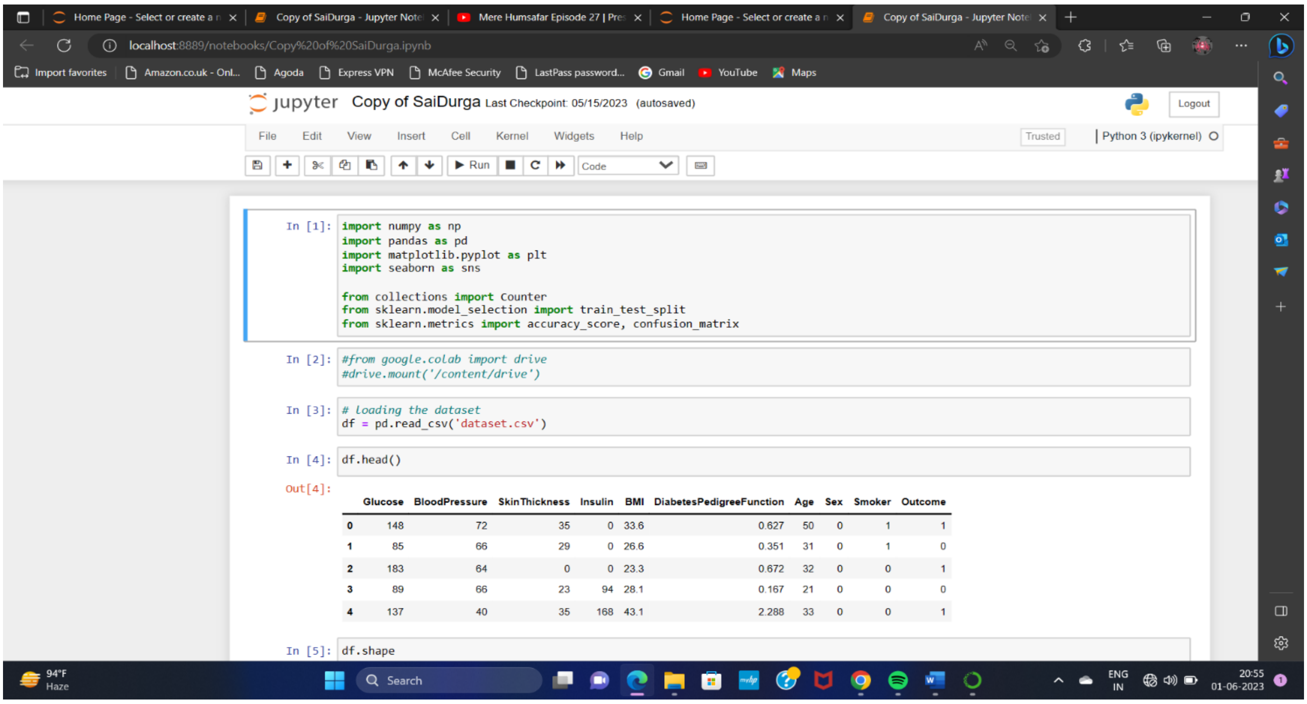
Open-source stages like Kubernetes are utilized to oversee and arrange holders, and robotize components of compartment the board like planning and scaling.

Predict Output

Using this module, we will upload test dataset and then classification model will predict output based on input data. In this module the user gives the different values as inputs by going with one of the algorithms SVM because of high accuracy than Logistic Regression. By this classification algorithm the data given by user is classified that user had a chance to get diabetes then it shows Diabetic, and the user had no chance to get diabetes then it shows Not Diabetic.

IJARST
International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

4. RESULTS AND DISCUSSION



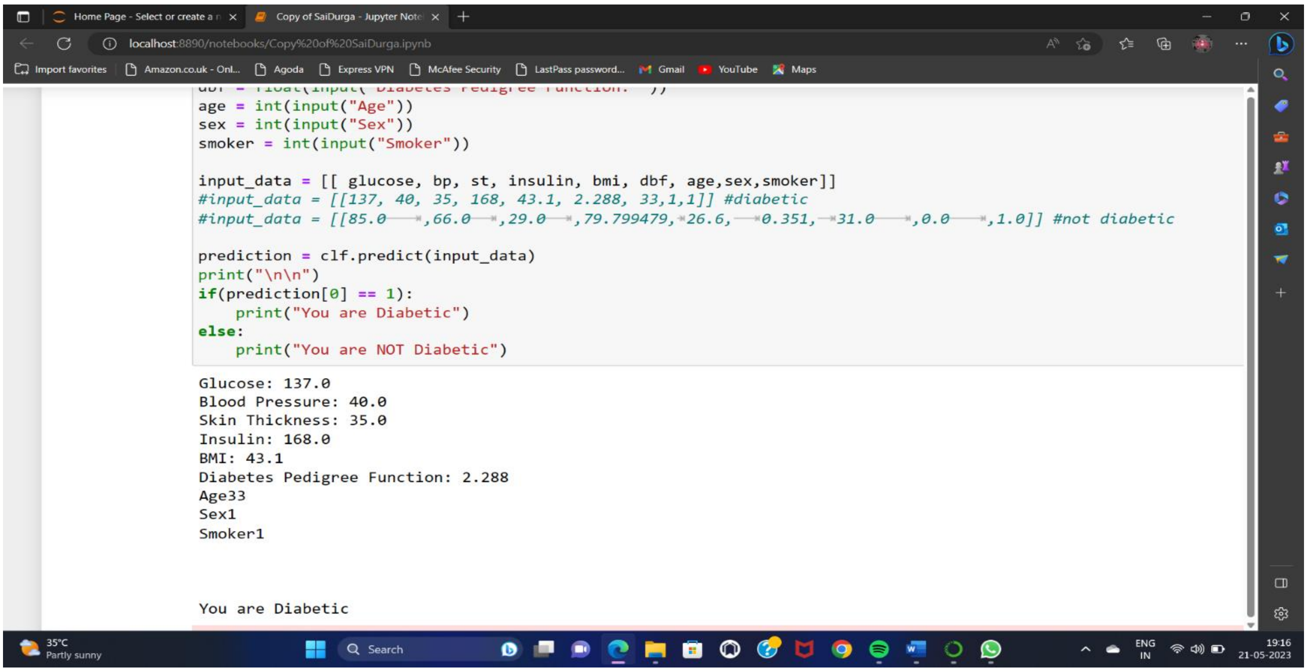
Copy of SaiDurga Last Checkpoint: 05/15/2023 (autosaved)

```
In [1]: import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt
        import seaborn as sns

        from collections import Counter
        from sklearn.model_selection import train_test_split
        from sklearn.metrics import accuracy_score, confusion_matrix

In [2]: #from google.colab import drive
        #drive.mount('/content/drive')

In [3]: # loading the dataset
        df = pd.read_csv('dataset.csv')

In [4]: df.head()

Out[4]:
```

	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Sex	Smoker	Outcome
0	148	72	35	0	33.6	0.627	50	0	1	1
1	85	66	29	0	26.6	0.351	31	0	1	0
2	183	64	0	0	23.3	0.672	32	0	0	1
3	89	66	23	94	28.1	0.167	21	0	0	0
4	137	40	35	168	43.1	2.288	33	0	0	1

```
In [5]: df.shape
```

Fig1: Loading the Dataset



```
dbf = float(input("Diabetes Pedigree Function: "))
age = int(input("Age"))
sex = int(input("Sex"))
smoker = int(input("Smoker"))

input_data = [[ glucose, bp, st, insulin, bmi, dbf, age,sex,smoker]]
#input_data = [[137, 40, 35, 168, 43.1, 2.288, 33,1,1]] #diabetic
#input_data = [[85.0    ,66.0    ,29.0    ,79.799479, 26.6,    0.351,  31.0     ,0.0     ,1.0]] #not diabetic

prediction = clf.predict(input_data)
print("\n\n")
if(prediction[0] == 1):
    print("You are Diabetic")
else:
    print("You are NOT Diabetic")
```

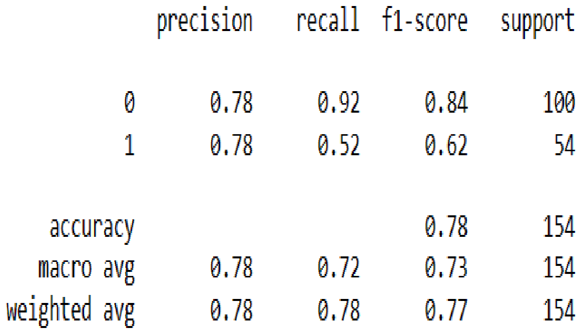
```
Glucose: 137.0
Blood Pressure: 40.0
Skin Thickness: 35.0
Insulin: 168.0
BMI: 43.1
Diabetes Pedigree Function: 2.288
Age33
Sex1
Smoker1



You are Diabetic
```

Fig 2: Displaying Output

The above diagram shows that code you provided is used to make a prediction on whether a person is diabetic or not based on input data.

4.1 CONFUSION MATRIX

An aid to visualizing a classification problem's results is a confusion matrix, which presents a table layout of the various outcomes of the prediction and results. It plots a table of all the predicted and actual values of a classifier. The matrix's cells each represent a unique combination of expected and observed values, making it simple to distinguish between true positives, true negatives, false positives, and false negatives. This data is essential for assessing a classification model's effectiveness. The matrix is commonly referred to as a confusion matrix, and it provides a comprehensive overview of the model's performance. It is often used to calculate various performance metrics, such as accuracy, precision, recall, and F1 score.



```
              precision    recall  f1-score   support

           0       0.78      0.92      0.84       100
           1       0.78      0.52      0.62        54

    accuracy                           0.78       154
   macro avg       0.78      0.72      0.73       154
weighted avg       0.78      0.78      0.77       154
```

Fig 3: Confusion Matrix

5. CONCLUSION

As the world transitions to a more economically based society, the goal is to stimulate each country's economy. The Type 2 diabetes in the person is predicted by the machine learning algorithm. By accurately predicting and preventing diseases like Type 2 diabetes, nations can reduce healthcare costs and boost the output of their workforce, which contributes to the development of a stronger economy. This demonstrates the importance of investing in the research and development of machine learning algorithms for use in healthcare applications.According to the clinical viewpoints, the SB-SVM model may be valuable likewise for the expectation of various obsessive conditions (e.g., cardiovascular and neurological infections). ML strategies that make use of EHR clinical data may be able to provide and anticipate early care strategies in addition to the resources provided by so-called predictive medicine, in which the point analysis of genetic and biological components represents the constitutive elements of forthcoming widespread implementation.

REFERENCES

1. Choi B.G., Rha S.-W., Kim S.W., Kang J.H., Park J.Y., Noh Y.-K. Machine learning for the prediction of new-onset diabetes mellitus during 5-year follow-up in non-diabetic patients with cardiovascular risks. *Yonsei Med. J.* 2019; 60:191–199.

2. Shaw J., Sicree R., Zimmet P. Global estimates of the prevalence of diabetes for 2010 and 2030. *Diabetes Res. Clin. Pract.* 2010; 87:4–14.

3. K. G. M. M. Alberti and P. Z. Zimmet, "Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: diagnosis and classification of diabetes mellitus provisional report of a WHO consultation." Diabetic Medicine, vol. 15, no. 7, pp. 539–553, 1998.

4. International Diabetes Federation, IDF Diabetes Atlas, 8th edn. Brussels, Belgium, 2017.

5. WHO et al., Global report on diabetes. World Health Organization, 2016.

6. B. Chaudhry, J. Wang, S. Wu, M. Maglione, W. Mojica, E. Roth, S. C. Morton, and P. G. Shekelle, "Systematic review: impact of health information technology on quality, efficiency, and costs of medical care," Annals of Internal Medicine, vol. 144, no. 10, pp. 742–752, 2006.

7. R. Kaushal, K. G. Shojania, and D. W. Bates, "Effects of computerized physician order entry and clinical decision support systems on medication safety: a systematic review," Archives of Internal Medicine, vol. 163, no. 12, pp. 1409–1416, 2003.

8. R. Amarasingham, L. Plantinga, M. Diener-West, D. J. Gaskin, and N. R. Powe, "Clinical information technologies and inpatient outcomes: a multiple hospital study," Archives of Internal Medicine, vol. 169, no. 2, pp. 108–114, 2009.

9. S. T. Parente and J. S. McCullough, "Health information technology and patient safety: evidence from panel data," Health Affairs, vol. 28, no. 2, pp. 357–360, 2009.

10. M. A. Figueiredo, R. D. Nowak, and S. J. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," IEEE Journal of Selected Topics in Signal Processing, vol. 1, no. 4, pp. 586–597, 2007.

11. J. Platt, "Probabilistic outputs for support vector machines and comparison to regularized likelihood methods," Advances in Large Margin Classifiers, pp. 61–74, 2000.

AUTHOR PROFILES

Ms.M.ANITHA completed her Master of Computer Applications and Masters of Technology. Currently working as an Assistant professor in the Department of Masters of Computer Applications in the SRK Institute of Technology, Enikepadu, Vijayawada, and NTR District. Her area of interest includes Machine Learning with Python and DBMS.

Ms. A. HARIKA VENKATA SIVA NAGA SAI DURGA is an MCA student in the Department of Computer Applications at SRK Institute of Technology, Enikepadu, Vijayawada, and NTR District. She had Completed Degree in B.Sc. (computers) from VijayaJyothi Degree College, Mangalagiri. Her areas of interest are DBMS, Java Script, and Machine Learning with Python.