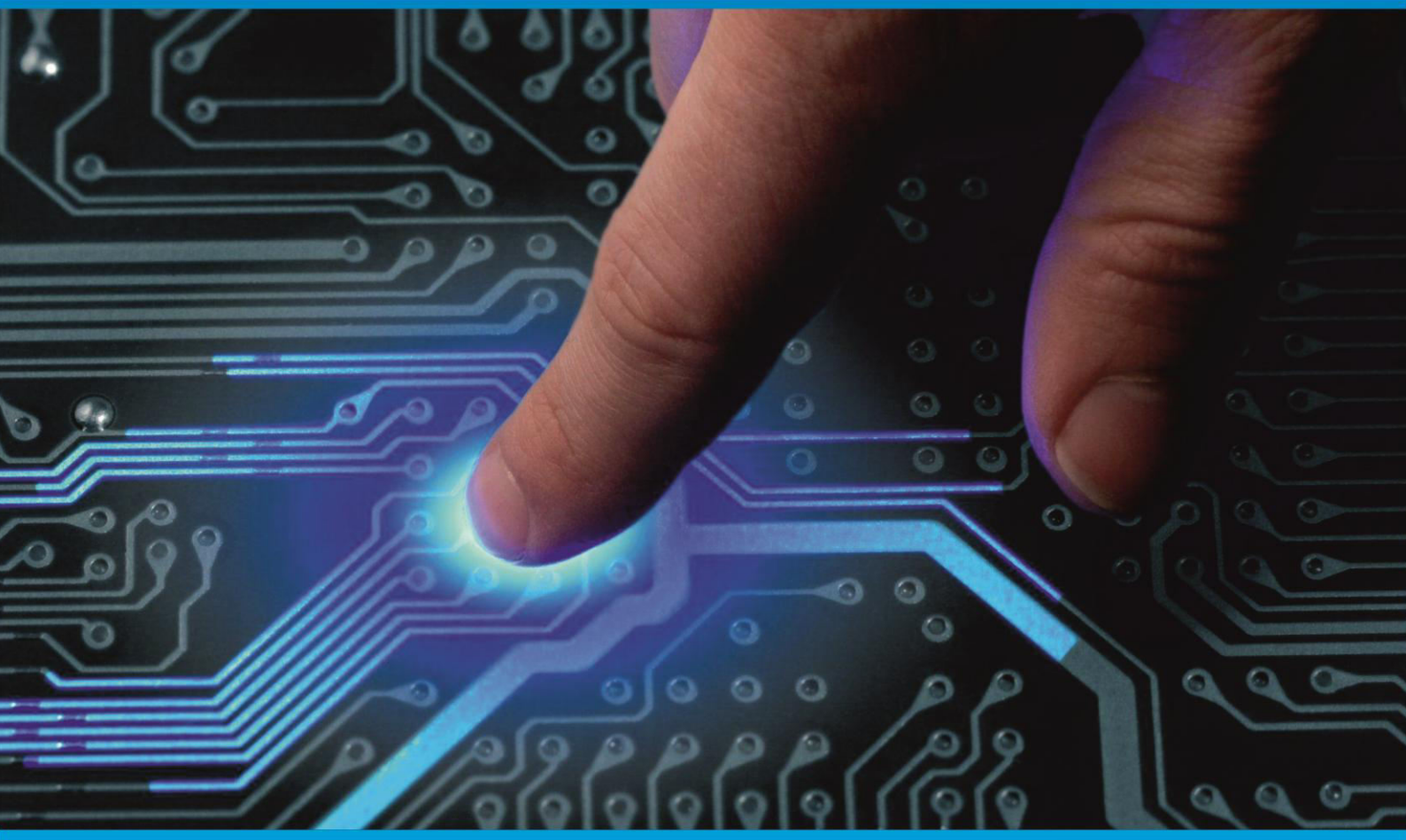
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 5, May 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.488

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 5, May 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0905238 |

Implementation on "Prediction of Heart Disease, Diabetes and Cancer using Machine Learning".

Omshree Ghodekar[1], Prafulla Bhaskar[2], Titiksha Shukla[3], Sakshi Shinde[4], Dr. M. A. Chaudhari[5]

UG Student, Dept. of Information Technology, Amrutvahini College of Engineering, Maharashtra, India[1]

UG Student, Dept. of Information Technology, Amrutvahini College of Engineering, Maharashtra, India[2]

UG Student, Dept. of Information Technology, Amrutvahini College of Engineering, Maharashtra, India[3]

UG Student, Dept. of Information Technology, Amrutvahini College of Engineering, Maharashtra, India[4]

Assist. Professor, Dept. of Information Technology, Amrutvahini College of Engineering, Maharashtra, India[5]

ABSTRACT: The system allows the user to make a use of algorithms to predict the risk of diabetes mellitus, cancer and heart disease in human body. The various classification models such as Decision Tree, K Nearest Neighbour, Support Vector Machine are used in this system. Among all the algorithms highest accuracy showing algorithms are used for each model in project. The dataset used is the Pima Indians Diabetes, Cancer and Heart disease Data Set, which has the information of patients, some of them have developing disease therefore, this project is aimed to create a mobile application for predicting a person's class whether present in of the diabetes,cancer and heart disease risk or not.

KEYWORDS: Machine Learning, AI, Disease Prediction, SVM, Random Forest, Decision Tree, Naïve Bayes, Ada Boost, Heart Disease, Cancer, Diabetes.

I. INTRODUCTION

Health-care information systems tend to capture data in databases for research and analysis in order to assist in making medical decisions. As a result, medical information systems in hospitals and medical institutions become larger and larger and the process of extracting useful information becomes more difficult. Traditional manual data analysis has become inefficient and methods for efficient computer-based analysis are needed. To this aim, many approaches to computerized data analysis have been considered and examined. Data mining represents a significant advance in the type of analytically tools. It has been proven that the benefits of introducing data mining into medical analysis are to increase diagnostic accuracy, to reduce costs and to save human resources. Cancer and diabetes are two destructive diseases in our society. Every year numerous people die out of cancer. The Agency for Healthcare Research and Quality (AHRQ) says that medical cost for cancer in the year 2011 in the United States was 88.7 billion dollars. And out of various types of cancer, breast cancer has been one of the significant types over the past years. Sometimes, breast cancer is detected at a stage when chances of survival are very low. Computer science can play some role to detect vulnerability of a cancer patient with medical data with the help of machine learning. By manipulating medical data, having attributes of cancer cell, a system can predict if the cancer is benign or malignant. If the cancer is in a benign stage, then taking appropriate measures can help the patient survive and can even heal them completely in some cases. Cancer and diabetes is another disease that kills people slowly. Cancer and diabetes has become prevalent almost all over the world

However, according to a study of Asian. Diabetic Prevention Organization, 60 percent of the whole worlds diabetic population is from Asia. So, Asian people are at high risk Existence of cancer and diabetes in a patient can be predicted by machine learning. So, in this research, cancer and diabetes is predicted as binary values like 1 or 0 meaning "YES" or "NO". The data set that is used for cancer and diabetes includes attributes of the patients feature that might lead to the existence of cancer and diabetes. Machine learns the attributes and then predicts in "YES" or "NO". The main objective of the research is to predict Cancer and Cancer and diabetes. For cancer it will predict the stage as

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 5, May 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0905238 |

"Malignant" or "Benign" and for cancer and diabetes it will predict as "YES" or "NO". The prediction is based on some of the state of the art machine learning algorithms. The project has another objective as to optimize the performances of these well-established machine learning algorithms. Some experiments will be performed to see if the algorithms can perform better on a different setup. Performance comparison is checked across different classifiers to understand how they behave with the same data set and how much time does each one take to build a classification model. One challenging prospect of this project is to achieve some techniques to apply curriculum learning on the data set.

II. RELATED WORK

Paper[1]"An Optimization Approach to Improve Classification Performance in Cancer and Diabetes Prediction" Some classification algorithms are experimented. Some optimization attempts are made to improve the algorithms performances. Detecting diseases like cancer and diabetes might be helpful for the patients as well as the doctors. From the doctor's perspective, they can help the patients to identify their next step by identifying the vulnerability of cancer or prevalence of diabetes in a patient. That is how the doctors may find a way to determine the patient's condition and also if someone is at a high risk of cancer the doctors can decide on the medication and a lifestyle to help them live a better life. Techniques used for Cancer and diabetes prediction. Abundant literature has been dedicated to the classifiers of data mining and tremendous progress has been made ranging from efficient and scalable algorithm for different datasets. Authors believe that mining research has substantially broadened the scope of data analysis and will have deep impact on mining methodologies and applications in the future. However, there are still some challenging research issues that need to be solved in searching using concept of various classifiers of data mining in the research of diabetes prediction.

Paper[2] "Type 2 diabetes mellitus prediction model based on data mining" WEKA toolkit and use the same Pima Indian Diabetes Dataset. realistic dataset provided by Dr. Schorling was used to test verify the model. K-means Algorithm and Logistic Regression: A novel model based on data mining techniques for predicting type 2 diabetes mellitus (T2DM). Based on a series of pre-processing procedures, the model is comprised of two parts, the improved K-means algorithm and the logistic regression algorithm. The Pima Indians Diabetes Dataset and the Waikato Environment for Knowledge Analysis toolkit were utilized to compare results with the results from other researchers.

Paper[3]An expert Personal Health System to monitor patients affected by Gestational Diabetes Mellitus. The graphs and charts provided in this study were of simple nature, as a future development, more elaborated charts representation of the data that puts physiological patterns in relation may allow the medical doctors to have a better understanding of the health of the patient, thus improving their ability to effectively modify the treatment plan. Temporal Abstraction with Data Mining: Haemodialysis patients might suffer from unhealthy care behaviours or longterm dialysis treatments and need to be hospitalized. If the hospitalization rate of a haemodialysis center is high, its service quality will be low. Therefore, decreasing hospitalization rate is a crucial problem for health care centres. This study combines temporal abstraction with data mining techniques for analysing dialysis patients biochemical data to develop a decision support system. The mined temporal patterns are helpful for clinicians to predict hospitalization of haemodialysis patients and to suggest immediate treatments to avoid hospitalization.

Paper[4] An improved electromagnetism-like mechanism algorithm and its application to the prediction of diabetes mellitus. In this study, 54 UCI datasets are used to evaluate the performance of various classification algorithms. These datasets are characterized according to data sizes, features, classes. Electromagnetism-like Mechanism: The use of artificial intelligence based data mining techniques for massive medical data classification and diagnosis has gained its popularity, whereas the effectiveness and efficiency by feature selection is worthy to further investigate. They Presented a novel method for feature selection with the use of opposite sign test (OST) as a local search for the electromagnetism-like mechanism (EM) algorithm, denoted as improved electromagnetism- like mechanism (IEM) algorithm. Nearest neighbour algorithm is served as a classifier for the wrapper method. The proposed IEM algorithm is compared with nine popular feature selection and classification methods. Forty-six datasets from the UCI repository and eight gene expression micro array datasets are collected for comprehensive evaluation. Non-parametric statistical tests are conducted to justify the performance of the methods in terms of classification accuracy and Kappa index. The results confirm that the proposed IEM method is superior to the common state-of- art methods.

Paper[5] Comparison of three data mining models for predicting diabetes or prediabetes by risk factors Author: Xue-Hui Meng Year : 2012. The characteristics of participants and Pearson Chi-square test results between two groups. The

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 5, May 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0905238 |

sensitivity analysis performed, logistic regression model, ANN model , C5.0 decision tree model, to determine the order detailed predictions produced from the training and testing datasets are presented in the form of confusion matrices. Pre-Diabetes Detection by Risk Factors: Purpose of this was to compare the performance of logistic regression, artificial neural networks (ANNs) and decision tree models for predicting diabetes or pre-diabetes using common risk factors. A standard questionnaire was administered to obtain information on demographic characteristics, family diabetes history, anthropomorphic measurements and lifestyle risk factors. Then "Improve the Performance of Cancer and Diabetes Detection by Using Novel Technique of Machine Learning" He developed three predictive models using 12 input variables and one output variable from the questionnaire information we evaluated the three models in terms of their accuracy, sensitivity and specificity. The logistic regression model achieved classification accuracy.

III. PROPOSED ALGORITHM

The process starts with data manipulation. Next, four models will be investigated for finding a prediction model. Then, accuracy of each model will be calculated and compared for seeking the best model. Detecting diseases like cancer, diabetes& heart attack might be helpful for the patients as well as the doctors. From the doctors' perspective, they can help the patients to identify their next step by identifying the vulnerability of cancer or prevalence of diabetes in a patient. The study ends up with creating aandroid application.

A. *System Design:*

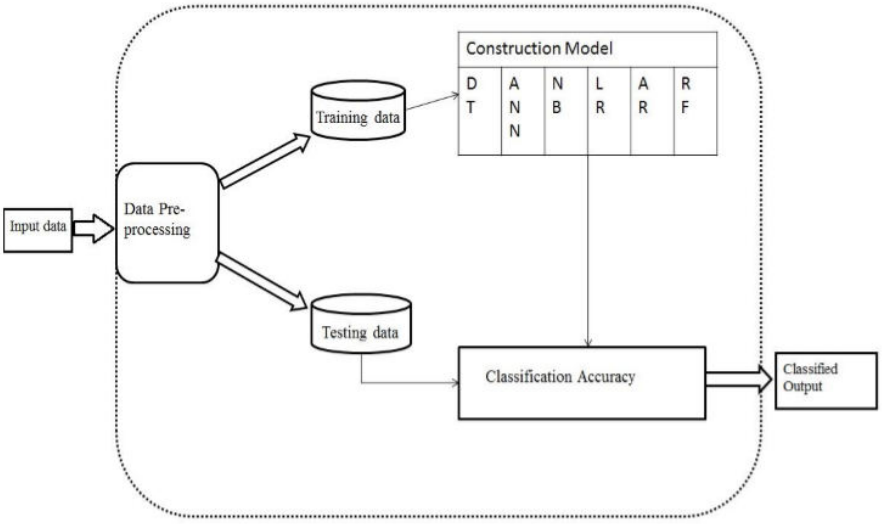


Fig 1: System Architecture

B. *Modules:*

1. Data Pre-processing:

First, each attribute's correlation to DM is analysed.

E.g.,number ofpregnancies is transformed into a nominal attribute. The value 0 indicateson-pregnant and 1 indicates pregnant. The complexity of the dataset wasreduced by this process. Second, some missing and incorrect values in thedataset due to errors are removed. For example, body mass index could not be0, which indicates that the real value was missing. To reduce the influence ofmeaningless values, the training data is used to replace all missing values.

2. Data Classification

The model consists of double-level algorithms. In the first level,authors usedthe improved k-means algorithm to remove incorrectly clustered data. The optimized dataset was used as input for next level. Then,they used the logistic regression algorithm to classify the remaining data.Improved k-means clustering algorithmcluster analysis aims at partitioning the observations into disparate. Clusters so that observations within the same cluster are more closely related toeach other than those assigned to different clusters. The K-means is one ofthe most popular cluster algorithms. It is a typical distance-based clusterAlgorithm, and the distance is used as a measure of similarity, i.e., the smallerdistance between objects shows the greater similarity.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 5, May 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0905238 |

C. *Basic Steps for Creating Model and Prediction:*

We used Scikit-learn library for creating the Models and calculating the accuracy. Also we used NumPy and pandas for data manipulation. We used Matplotlib library for Data visualization. Steps for creating model is as follows:

1. Import the libraries e.gNumPy, pandas, Sklearn, matplotlib.
2. Load the dataset which is in csv file using **pandas.read_csv()** method.
3. Identify the features and classes from the dataset.
4. Then we replace the missing values with either 0's or replace them with the mean using method **pandas.fillna()**. Eliminating missing values results into better accuracy.
5. As machine learning algorithms take the input in the form of numerical data we need to preprocess the data for that we can use some scalars like **StandardScaler() or MinMaxScaler().** These Scalers are available in **sklearn.preprocessing** package. We'll use **scaler.fit_transform()** method for preprocessing.
6. Then we'll split the data into training and testing. For training we used 80% data and for testing we used 20% data. We used **train_test_split()** method from **sklearn.model_selection** package.
7. Now we'll define the models. There are various models available in the **sklearn** library like Random Forest, Decision Tree, SVM, etc. then we'll fit the data to model using **model.fit()** method.
8. After defining and fitting the data to model we calculated accuracy and the confusion matrix for each model. It is best practice to do this so that you can know which model fits the data perfectly and can predict the data with better accuracy. For Calculating the accuracy we used **accuracy_score()** method from **sklearn.metrics**package we'll provide the **y_test**and **best_prediction**to the **accuracy_score()**method. To calculate the confusion matrix we'll use **confusion_matrix()** method from **sklearn.metrics**package.
9. For final prediction we'll use **predict()** method of the model.

IV. PSEUDO CODE

A. For Heart Disease &Cancer Prediction KNN:

K-nearest neighbors (KNN) algorithm is a type of supervised ML algorithm which can be used for both classification as well as regression predictive problems. However, it is mainly used for classification predictive problems in industry.

Step-1: Select the number K of the neighbours–

Step-2: Calculate the Euclidean distance of K number of neighbours–

Step-3: Take the K nearest neighbours as per the calculated Euclideandistance.

Step-4: Among these k neighbours, count the number of the Data pointsin each category.

B. For Diabetes Prediction Decision Tree:

1. Check if algorithm satisfies termination criteria
2. Computer information-theoretic criteria for all attributes
3. Choose best attribute according to the information-theoretic criteria
4. Create a decision node based on the best attribute in step
5. Induce (i.e. split) the dataset based on newly created decision node in step 4
6. For all sub-dataset in step 5, call C4.5 algorithm to get a sub-tree (recursive call)
7. Attach the tree obtained in step 6 to the decision node in step 4
8. Return tree

V. RESULTS & DISCUSSIONS

A. Analysis

1) For Heart Disease:

For Heart Disease prediction we selected 5 Machine Learning Algorithms viz. Decision Tree, KNN, Naïve Bayes, Random Forest, SVM Linear. Each model was trained on the data and we calculated accuracy for each model. Accuracy of each model is listed below:

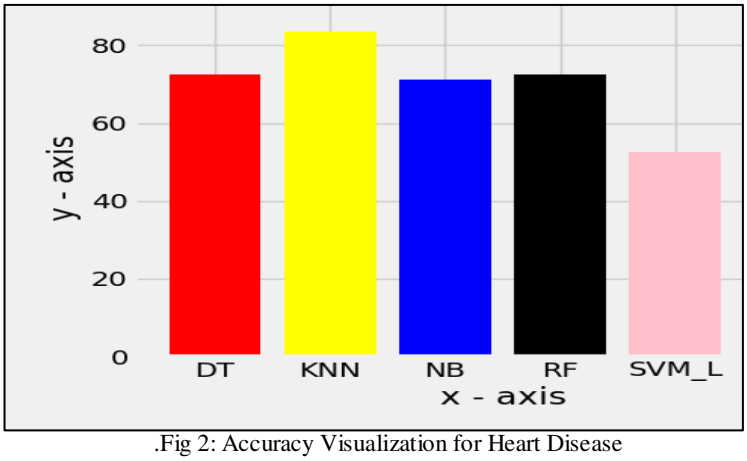
- DECISION TREE - 72.36%
- KNN- 83.60%
- NAIVE BAYES- 63.25%

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 5, May 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0905238 |

- RANDOM FOREST - 67.10%
- SVM-L - 52.63%

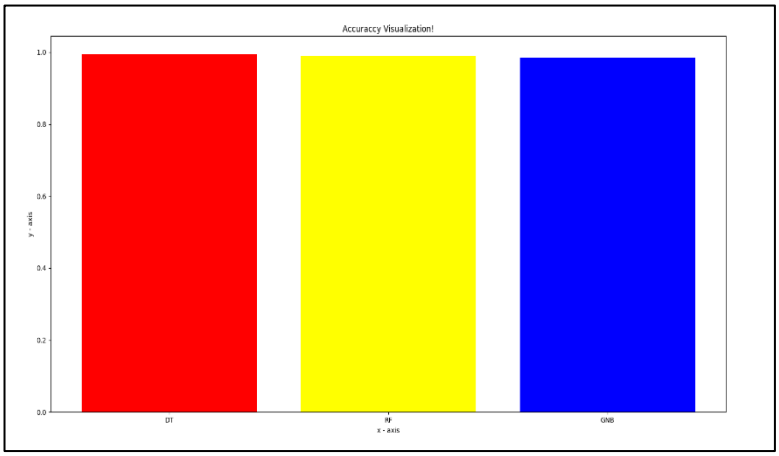


.Fig 2: Accuracy Visualization for Heart Disease

KNN gives highest accuracy, hence it is used for further predictionof heart disease.

2) For Diabetes Disease

Total records given as input for prediction: 760
247 patient's record gave outcome as 0
253 patient's record gave outcome as 1
269 patient's record gave outcome as 2
For Diabetes prediction we used 3 Algorithm and created model for them they are Decision Tree, Random Forest and Naïve Bayes. Accuracy of each model is listed below:



- DECISION TREE - 0.9947
- RANDOM FOREST - 0.9895
- NAIVE BAYES - 0.98

Fig 3: Accuracy Visualization for diabetes

Decision tree shows highest accuracy hence it is used for furtherprediction of Diabetes.

3) For Cancer Detection

For Cancer Prediction we used 3 Algorithms and created models for them they're KNN, SVM and Random Forest. The accuracy of each model is below:

- KNN - 96%
- SVM - 95%

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 5, May 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0905238 |

- RM - 95.2%

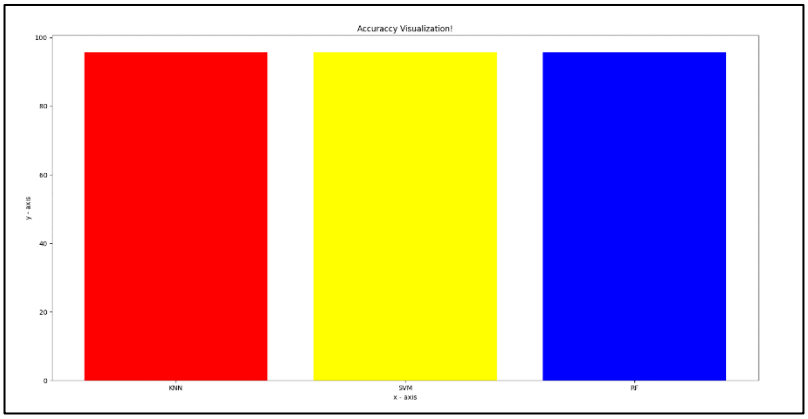


Fig 4: Accuracy Visualization for cancer

KNN shows highest accuracy hence KNN is used for further prediction of cancer.

VI. CONCLUSION AND FUTURE WORK

The proper filters were utilized to improve the validity and rationality of thedataset. The proposed model that consisted of both cluster and class method ensuredthe enhancement of prediction accuracy .From experimental results, we find that KNN gives highest accuracy for heart disease and cancer , hence it is used for prediction of same. While DECISION TREE showshighest accuracy for prediction of Diabetes . It assures less time consuming and maximum retention of original data. Themain problems solved are improving accuracy of prediction model and making themodel to adapt to different datasets.

REFERENCES

1. Mustakim Al Helal, Atiqul Islam Chowdhury, Ashraful Islam, EshtiakAhmed,Md. Swakshar Mahmud, Sabrina Hossain, “An Optimization Approach to Im-prove Classification Performance in Cancer and Diabetes Prediction”, Interna-tional Conference on Electrical, Computer and Communication Engineering(ECCE), 7-9 February, 2019(Base Paper)
2. Stefano Bromuri, SerbanPuricel, Rene Schumann, Johannes Krampf, JuanRuiz and Michael Schumacher, “An expert Personal Health System to mon-itor patients affected by Gestational Diabetes Mellitus: A feasibility study”,Journal of Ambient Intelligence and Smart Environments 8(2016) 219–237.
3. Gyorgy J. Simon, Pedro J. Caraballo, Terry M. Therneau, Steven S. Cha, M.Regina Castro and Peter W. Li, “Extending Association Rule SummarizationTechniques to Assess Risk of Diabetes Mellitus”, IEEE TRANSACTIONSON KNOWLEDGE AND DATA ENGINEERING.
4. Han Wu, Shengqi Yang, Zhangqin Huang, Jian He, Xiaoyi Wang, “Type 2 dia-betes mellitus prediction model based on data mining”, Informatics in MedicineUnlocked 10 (2018) 100–107.
5. Xue-Hui Meng a, Yi-Xiang Huang a, Dong-Ping Rao b, Qiu Zhang a, QingLiu b, “Comparison of three data mining models for predicting diabetes orprediabetes by risk factors”, Kaohsiung Journal of Medical Sciences (2013)29,93e99.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.488

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर
NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details