OPEN ACCESS INTERNATIONAL JOURNAL OF SCIENCE & ENGINEERING

GENERAL DISEASE PREDICTION BASED ON SYMPTOMSPROVIDED BY PATIENT

Nishant Yede [1], Ritik Koul [2], Chetan Harde [3], Kumar Gaurav [4], Prof. C.S.Pagar [5]

[1234] *B. E Students, IT Department, SKNSITS, Lonavala, India*

[5]*Assistant Professor, IT Department, SKNSITS, Lonavala, India*

nishant.yede24@gmail.com [1], *ritikkoul777gmail.com* [2], *chetanharde02@gmail.com* [3], *19kumarg@gmail.com* [4], *csp.sknsits@sinhgad.edu* [5]

Abstract: **With big data growth in biomedical and healthcare communities, accurate analysis of medical data benefits early disease detection, patient care, and community services. However, the analysis accuracy is reduced when the quality of medical data is incomplete. Moreover, different regions exhibit unique characteristics of certain regional diseases, which may weaken theprediction of disease outbreaks. In this paper, we streamline machine learning algorithms for effective prediction of chronic disease outbreak in disease-frequent communities. The application of machine learning in the field of medical diagnosis is increasing gradually. This can be contributed primarily to the improvement in the classification and recognition systems used in disease diagnosis which is able to provide data that aids medical experts in early detection of fatal diseases and therefore, increase the survival rate of patients significantly.**

The results of the study strengthen the idea of the application of machine learning in early detection of diseases. Compared toseveral typical calculating algorithms, the scheming accuracy of our proposed algorithm reaches 94.8% with a regular speed which is quicker than that of the unimodel disease risk prediction algorithm and produces report.

Keywords: *Machine Learning, Django, Navie Bayes, HealthCare Symptoms.*

I INTRODUCTION

Due to big data progress in biomedical and healthcare communities, accurate study of medical data benefits early disease recognition, patient care and community services. When the quality of medical data is incomplete the exactness of study is reduced. Moreover, different regions exhibit unique appearances of certain regional diseases, which may results in weakening the prediction of disease outbreaks.

In this project, it bid a Machine learning Decision tree map, Navie Bayes, Random forest algorithm by using structured and unstructured data from hospital. It also uses Machine learning algorithm for partitioning the data. To the highest of gen, none of the current work attentive on together data types in the zone of remedial big data analytics. Compared to several typical calculating algorithms, the scheming accuracyof our proposed algorithm reaches 94.8% with an regular speed which is quicker than that of the unimodel disease risk prediction algorithm and produces report As per the Centres for Medicare and Medicaid services, 50% of Americans have multiple chronic diseases with a total US health care expenditure in 2016 to be about $3.3trillion, which amounts to $10,348 per person in the US. With the growth in medical data collecting electronic health records (EHR) is increasingly convenient Besides, first presented a bio- inspired high-performance Heterogeneous vehicular telematics paradigm, such that the Collection of mobile users' health-related real-time big data can be achieved with the deployment of advanced heterogeneous vehicular networks. Chen et al. proposed a healthcare system using smart clothing for sustainable health monitoring. Qiu *et al.* had thoroughly studied the heterogeneous systems and achieved the best results for cost minimization on tree and simple path cases for heterogeneous systems. Patients' statistical information, test results and disease history are recorded in the EHR, enabling us to identify potential data-centric solutions to reduce the costs of medical case studies.

With the development of big data analytics technology, more attention has been paid to disease prediction from the perspective of big data analysis, various researches have been conducted by selecting the characteristics automatically from a large number of data to improve the accuracy of risk classification rather than the previously selected characteristics. However, those existing work mostly considered structured data.

II LITERATURE SURVEY

- M. Chen, Y. Hao, K. Hwang, L. Wang and L. Wang, developed and propose a new convolutional neural network based multimodal disease risk prediction (CNN-MDRP) algorithm using structured and unstructured data from hospital. To the best of our knowledge, none of the existing work focused on both data types in the area of medical big data analytics. Compared to several typical prediction algorithms, the prediction accuracy of our proposed algorithm reaches 94.8% with a convergence speed which is faster than that of the CNN-based unimodel disease risk prediction (CNN-UDRP) algorithm.[1]
- J. Gao, L. Tian, J. Wang, Y. Chen, B. Song and X. Hu, this study confirm the application of machine learning algorithms in prediction and early detection of diseases. To our best understanding, the model built according to the proposed method exhibits better accuracy than the existing ones .The prediction accuracy of our proposed method reaches 87.1% in Heart Disease detection using Logistic Regression, 85.71% in Diabetes prediction using Support Vector Machine (linear kernel) and98.57% using AdaBoost classifier for Breast Cancer detection. The future scope and improvement of the project involve automation of the steps such as data mugging, feature selection and model fitting for best prediction accuracy. Use of pipeline structure for data pre-processing could further help in achieving improved results.[2]
- A. N. Repaka, S. D. Ravikanti and R. G. Franklin, proposed data collection is carried out using numerous sources that are primary factors responsible for any sort of heart disease and thereby using a structure the database is constructed. The research focuses on establishing SHDP (Smart Heart Disease Prediction that takes into consideration the approach of NB (Naive Bayesian) classification and AES (Advanced Encryption Standard) algorithm for resolving the issue of heart disease prediction. its revealed that in regard to accuracy, the prevailing technique surpasses the Naive Bayes by yielding an accuracy of 89.77%in spite of reducing the attributes. AES yields in high security performance evaluation in comparison to PHEA (Parallel Homomorphic Encryption Algorithm).[3]
- P. S. Kohli and S. Arora, for heart disease prediction SVM,Naive Bayes and Decision tree has been applied with and without using PCA on the dataset. We used PCA to reduce the number of attributes. After reducing the size of the dataset, SVM outperforms Naive Bayes and Decision tree. SVM can further be used to predict heart disease. A GUI desktop application can be built using SVM and this dataset to predict the possibility of cardiovascular disease in a patient and for diabetes data prediction, the main aim of this paper is to predict diabetes disease using WEKA data mining tool. Our algorithms were implemented using WEKA data mining technique to analyse algorithm accuracy which was obtained after running these algorithms in the output window. These algorithms compare classifier accuracy toeach other on the basis of correctly classified instances, time taken to build model, mean absolute error and ROC Area. So, using above all observations, we can conclude that Maximum ROC Area means excellent predictions performance as compared to other algorithms[4]
- Arezou Koohi and Houman Homayoun ,Jie Xu , Mahdi Orooji develops Big data and Machine Learning have changed the health care research in recent years. Data generated from Electronic Health Records (EHRs) and other clinical sources now can be used further to help the patients. By applying Big Data Analytics (BDA) into health care data, it is possible to predict the outcome or the effects of drugs or risk of developing disease onhuman body. Several machine learning algorithms such as clustering, classification are used to analyze health care data. In this article, a framework is proposed using C-means Clustering for Biomedical Engineering applications. The framework can be used to help both the clinicians and the patients. For example, using this framework, a clinician can make a decision to prescribe suitable drug to a particular patient. In order to develop this framework, data has been collected from UCI machine learning repository. The data then analyzed using a well-known big data framework Hadoop. [5]

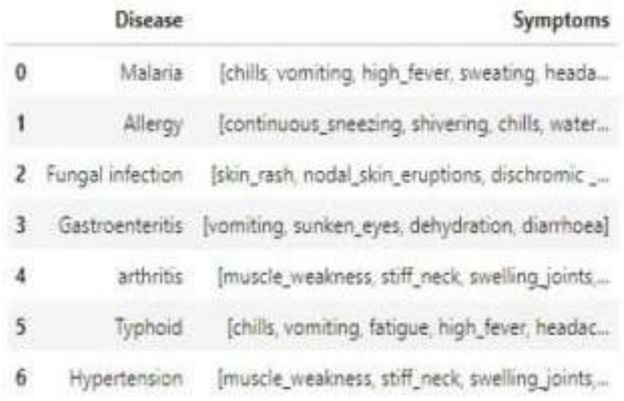
III RELATED WORK

1] Data collection

Data collection has been done from the internet to identify the disease here the real symptoms of the disease are collected i.e. no dummy values are entered.

The symptoms of the disease are collected from kaggle.com and different health related websites. This csv file contain 5000 rows of record of the patients with their symptoms (132 types of different symptoms) and their Corresponding disease (40 class of general disease).

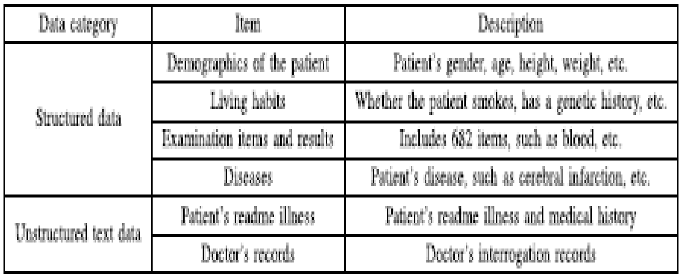
Some rows of disease with their corresponding symptoms inthe dataset are -



	Disease	Symptoms
0	Malaria	[chills, vomiting, high_fever, sweating, heada...
1	Allergy	[continuous_sneezing, shivering, chills, water...
2	Fungal infection	[skin_rash, nodal_skin_eruptions, dischromic _...
3	Gastroenteritis	[vomiting, sunken_eyes, dehydration, diarrhoea]
4	arthritis	[muscle_weakness, stiff_neck, swelling_joints,...
5	Typhoid	[chills, vomiting, fatigue, high_fever, headac...
6	Hypertension	[muscle_weakness, stiff_neck, swelling_joints,...

2. Data Classification

The Model Consists Of Double-Level Algorithms. In The First Level, Authors Used The Improved K-Means AlgorithmTo 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.



Data category	Item	Description
Structured data	Demographics of the patient	Patient's gender, age, height, weight, etc.
	Living habits	Whether the patient smokes, has a genetic history, etc.
	Examination items and results	Includes 682 items, such as blood, etc.
	Diseases	Patient's disease, such as cerebral infarction, etc.
Unstructured text data	Patient's readme illness	Patient's readme illness and medical history
	Doctor's records	Doctor's interrogation records

IV PROPOSED WORK

The proposed system has various users namely Administrator, Doctor, Analyst/Researcher. The role of the administrator is to add or remove users. The Doctor's role includes naming the disease and their symptoms in database. The role of the analyst is to choose the parameters for the analysis and apply K-means & SVM algorithm to the data. The parameters can be in the form of dates, gender or age. Once the analyst chooses required parameter, he/she can select the representation method in which the desired output will be displayed. Proposed system will be able to handle patient queries from start to end. It will respond with appropriate answers to queries when asked. The device would be lightweight with as many questions and responses aspossible.

- The key purpose of our method is that if consumers do not have experience of the medical profession and want to know about their health conditions, they will quickly find it without a technological or medical person.
- Machine learning is the core principle in which the system provides more precise forecasts.
- The Naive Bayes algorithm is more reliable on a similar situation than the medical field estimation.

V METHODOLOGY

In our system, there are three Modules: Admin, User (Patient), Doctor. Every new user has to get registered through admin. After successful registration user needs to enroll first before login. Users will need to enroll only once.

The disease prediction system have 3 users such as doctor, patient and admin.

- Each user of the system are authenticated by the system.
- There is a role based access to the system.
- The system allows the patient to give symptoms and according to those symptoms the system will predict a disease.
- The system suggests doctors for predicted diseases.
- The system allows online consultation for patients.
- The system helps the patients to consult the doctor at their convenience by sitting at home.

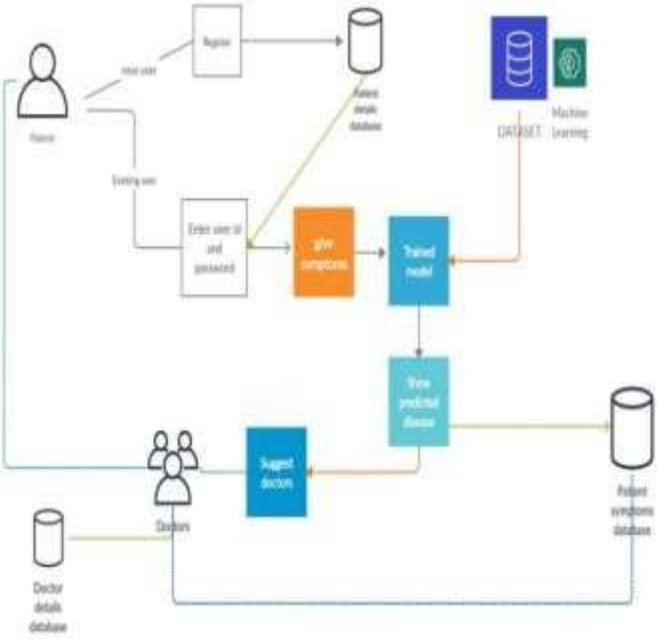


Figure 1:- System Architecture

VI ALGORITHM AND MATHEMATICAL MODEL

a) Naive Bayes classifier

The supervised machine learning method of classification is represented by Naive Bayes algorithm. It uses a probabilistic model by determining probabilities of the outcomes/outputs. It is used in analytical and predictive problems. Naive Bayes is robust to noise in input dataset. An implementation of Naive Bayes has been illustrated in Figure 2.

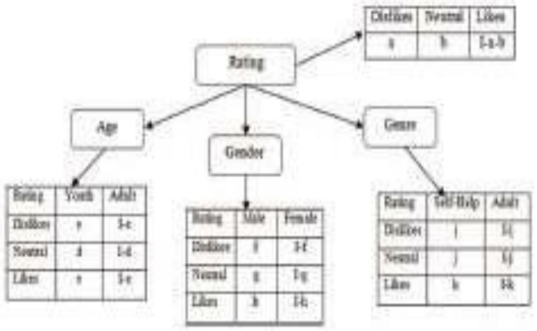


Fig 2:- Navie Bayes

b) Decision Tree

The decision tree learning is like as decision tree algorithm which uses maps input about an item to output of the item. The tree models with finite classes of output are called classification trees. In these tree structures leaves shows class labels and branches shows relation between attributes that the results in those class labels of the system.

Decision trees with continuous output classes are called regression trees. In data mining, a decision tree can be an input for decision making. An example of decision tree is demonstrated in Figure 3.

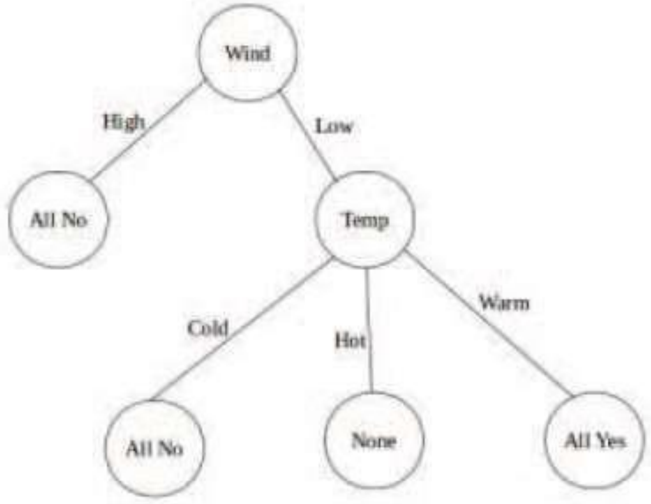


Fig 3:- Decision Tree

c) Random Forest Algorithm

Random Forest algorithm developed from trees algorithm and bagging algorithm is modelled. The developed the algorithm found that it can potentially improve classification accuracy. It is also work well with a data set with large number of input variables. The algorithm is started by creating a combination of trees which each will vote for a class as shown in Fig. The figure presents how to model the Random Forest.

Suppose that there are N data and M input variables in a data set where the real data used in this paper compose of data and input variables. Let k be the number of sampling groups, ni and mi be number of data and variables in group i where i is equal to 1, 2, ... and k.

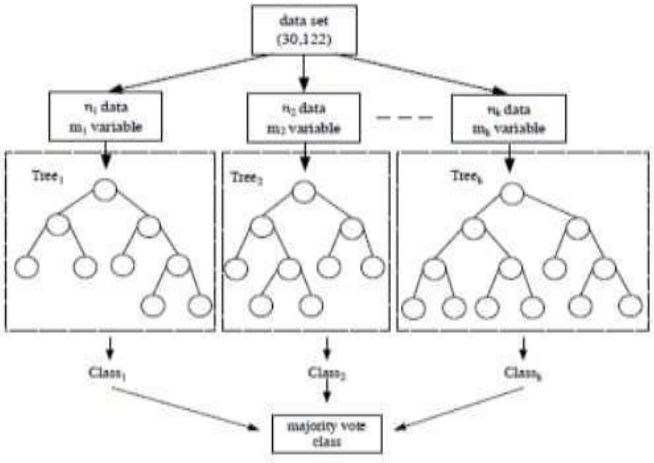


Fig 4:- Random Forest Model

2] MATHEMATICAL MODEL:-

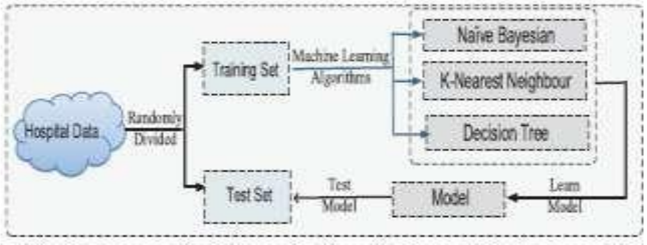
Let S be the whole System,
Set S = I, P, O Where,

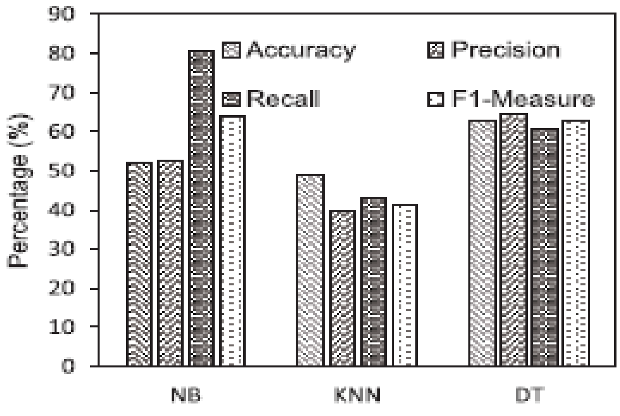
Input (I) represented as: I = {I0, I1, I2, I3, I4}
I0 = Patient Registration Details
I1 = Patient Login
I2 = upload patient health record
I3 = doctor check patient medical record
I4 = give prescription

Process (P) represented as: P = {P0, P1, P2, P3, P4}
P0 = Login by Patient-side
P1 = Login by doctor-side
P2 = Approval of login P3
= SVM, CNN
P4 = Logistic Regression & Random Forest

Output (O) represented as: O = {O0, O1, O2, O3, O4, O5}O0 = show Patient details
O1 = receiver id
O2 = Predict Disease
O3 = Predict Accuracy
O4 = Get Prescription
O5 = view details

VII RESULT ANALYSIS





(a)

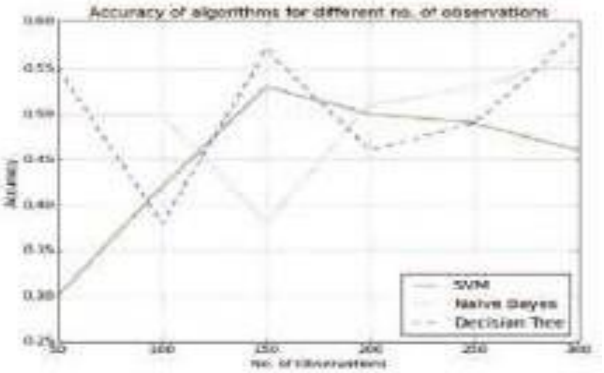
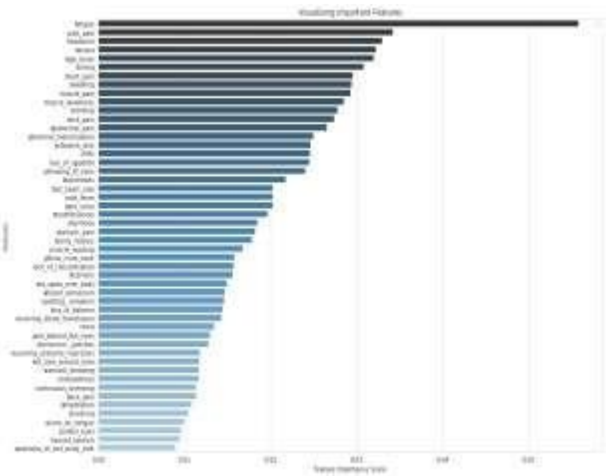


Fig:- Number of observations vs Accuracy



VIII CONCLUSION AND FUTURE SCOPE

The results of this study confirm the application of machine learning algorithms in prediction and early detection of diseases. To our best understanding, the model built according to the proposed method exhibits better accuracy than the existing ones. Compared to several typical calculating algorithms, the scheming accuracy of our proposed algorithm reaches 94.8% with an regular speed which is quicker than that of the unimodal disease risk prediction algorithm and produces report.

The Further work will mainly focus on the Medical Assistance and proper Medication to the patients as soon as possible soo as to build the best infrastructure and quickeasiest way in the Medical sectors.

ACKNOWLEDGMENT

This work can be used in Medical field and on basic ground level in all parts of world. Thus strengthening the Medical and Healthcare Industries. Authors are thankful to Faculty of Engineering and Technology (FET), Savitribai Phule Pune University, Pune for providing the facility to carry out the research work.

REFERENCES

[1] M. Chen, Y. Hao, K. Hwang, L. Wang and L. Wang, **"Disease Prediction by Machine Over Learning Over Big Data From Healthcare Communities,"** in IEEE Access, vol. 5, pp. 8869-8879, 2017, doi: 10.1109/ACCESS.2017.2694446.

[2] J. Gao, L. Tian, J. Wang, Y. Chen, B. Song and X. Hu, **"Similar Disease Prediction With Heterogeneous Disease Information Networks,"** in IEEE Transactions on Nano Bioscience, vol. 19, no. 3, pp. 571-578, July 2020, doi: 10.1109/TNB.2020.2994983.

[3] A. N. Repaka, S. D. Ravikanti and R. G. Franklin, "**Design And Implementing Heart Disease Prediction Using Naives Bayesian,"** 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2019, pp. 292-297, doi: 10.1109/ICOEI.2019.8862604.

[4] P. S. Kohli and S. Arora, **"Application of Machine Learning in Disease Prediction,"** 2018 4th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, 2018, pp. 1-4, doi: 10.1109/CCAA.2018.8777449.

[5] D. Le, **"Disease phenotype similarity improves the prediction of novel disease-associated microRNAs,"** 2015 2nd National Foundation for Science and Technology Development Conference on Information and Computer Science (NICS), Ho Chi Minh City, 2015, pp. 76-81, doi: 10.1109/NICS.2015.7302226.

[6] B. D. Kanchan and M. M. Kishor, **"Study of machine learning algorithms for special disease prediction using principal of component analysis,"** 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC), Jalgaon, 2016,pp.510,doi:10.1109/ICGTSPICC.2016.7955260.

[7] D. E. O'Leary**,"Ethics for Big Data and Analytics,"** in IEEEIntelligent Systems, vol. 31, no. 4, pp. 81-84, July-Aug. 2016, doi: 10.1109/MIS.2016.70.

[8] D. Dahiwade, G. Patle and E. Meshram**, "Designing Disease Prediction Model Using Machine Learning Approach,"** 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), 2019, pp. 1211-1215, doi: 10.1109/ICCMC.2019.8819782.

[9] Z. Sun, H. Yin, H. Chen, T. Chen, L. Cui and F. Yang, **"Disease Prediction via Graph Neural Networks,"** in IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 3, pp. 818-826, March 2021, doi: 10.1109/JBHI.2020.3004143.

[10] P. B. Jensen, L. J. Jensen and S. Brunak, **"Mining electronic health records: Towards better research applications and clinical care"**, *Nature Rev. Genetics*, vol. 13, no. 6, pp. 395-405, 2012.

****

e- National Conference

On

Advances in Modern Technologies of Multidisciplinary Research in Engineering Field (AIMTMREF)

[20th -21st May, 2021]

In association with ISTE , IETE and CSI

Address for Correspondence SKN Sinhgad Institute of Technology and Science Lonavala, Pune. 410 401, MS, India.

Website: www.sinhgad.edu

****