iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** IV **Month of publication:** April 2022

DOI: https://doi.org/10.22214/ijraset.2022.41588

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue IV Apr 2022- Available at www.ijraset.com

Common Ailments Possibility Using Machine Learning

Mohammed Sameer[1], Omjee[2], Raghav Gupta[3], Ritik Tyagi[4], Annu Mishra[5]
[1, 2, 3, 4, 5]*Inderprastha Engineering College*

Abstract: ***Machine learning in healthcare helps humans to process large and complex medical datasets and then analyze them into clinical insights which can help physicians in providing better medical care. Therefore, machine learning, when implemented in the medical field can lead to increased patient satisfaction. In this research, we will try to implement the functionalities of machine learning in healthcare in a single system. Health care can be made smart with the help of machine learning. Many cases can occur when the early diagnosis of an ailment is not within reach, So, their ailment prediction cannot be effectively implemented. As widely said "Prevention is better than cure", prediction of diseases would lead to early prevention of occurrence of disease. Medical Staff are often overworked in the medical field and hence the diagnosis becomes prone to human errors and negligence. Patients should be given treatment and diagnosis that are accurate and precise. Mistreatment may result in worsening the condition of the patient and hence the need for precise diagnosis. Therefore, the application of machine learning in disease prediction is considered in this paper as the best practice to facilitate a better healthcare system and provide better treatment to a patient as soon as possible. This paper majorly focuses on the development of a web app that would work on symptoms collected from the user and medical data and store it in the system. This data then will be analyzed using different machine learning algorithms to deliver results with maximum accuracy.***
Keywords: Machine Learning, Random forest, Support Vector Machine, Supervised learning.

I. INTRODUCTION

Medicine and healthcare are some of the most crucial parts of the economy and human life. There is a tremendous amount of change in the world we are living in now. Nowadays, hospitals are well equipped with monitoring and other data collection devices resulting in enormous data which are collected continuously through health examination and medical treatment.

In this situation, where everything has changed so much, the doctors and nurses are putting up maximum efforts to save people's lives even if they have to risk their own. In remote areas where there is a lack of medical facilities, some virtual assistance in case of emergency can play a major role. Machines are always considered better than humans as, without any human error, they can perform tasks more efficiently and with a consistent level of accuracy. A disease predictor can be called a virtual doctor, which can predict the disease of any patient without any human error. Also, in situations like COVID-19, a disease predictor can be a blessing as it can identify a person's disease without any physical contact. Disease Prediction using Machine learning is a system that predicts the disease based on the symptoms provided by the user. It is a system that also provides the user with a specialized doctor for the disease predicted. It is a system that provides the user the tips and tricks to maintain the health system of the user.

The purpose of making this project is to predict the accurate disease using the symptoms provided by the user. Using this information, we will compare with our previous datasets of the patients and predict the disease of the patient he/she has been through. If this prediction is done at the early stages of the disease can be cured and in general, this prediction system can also be very useful in the health industry. If the health industry adopts this project, then the work of the medical staff can be reduced and they can easily diagnose the disease of the patient. The general purpose of this research is to provide a prediction for the various and generally occurring diseases that when unchecked and sometimes ignored can turn into dangerous diseases and cause a lot of problems to the patient and as well as their family members. This system will predict the most possible ailment based on the symptoms provided. So, with the help of all these algorithms, techniques, and methodologies we have done this project with hopes to help the people who are in need.

II. LITERATURE REVIEW

Iwendi C [et.al] in 2020 proposed a fine-tuned Random Forest model boosted by the AdaBoost algorithm. The model uses the COVID-19 patient's geographical, travel, health, and demographic data to predict the severity of the case and the possible outcome, recovery, or death. The model has an accuracy of 94% and an F1 score of 0.86 on the dataset used. The data analysis reveals a positive correlation between patients' gender and deaths and indicates that most patients are aged between 20 and 70 years.

Keniya [et.al] in their research done in 2020 used different machine learning models to examine the prediction of disease for available input datasets. The authors used 11 different ML models for the prediction. Out of the 11 models, they managed to get 50 % or above accuracy for 6 models. As shown in the table, The Highest accuracy of a few of the models are SVM, Random Forest, and naïve Bayes. The Table 2 shows the comparison of the models they used along with their respective accuracies.

Method	Model Used	Maximum Accuracy(%)
Mir et al. [1]	Naive Bayes, SVM Random Forest and Simple CART	79.13
Vijayarani et al. [3]	SVM	79.66
Sriram et al. [5]	Random Forest	90.26

Table 1: Comparison of methodologies reported in the existing literature

Disease	Accuracy	Size of Original Feature Vector	Size of reduced Feature Vector
Heart	75.4%	13	10
Liver	75.9%	8	6
Diabetes	78.6	10	8

Table 2 Accuracy on different datasets of the same model with feature selection

Anant Agrawal [et.al] in 2018 proposed a hybrid machine learning model comprising of genetic algorithm and support vector machine. They have tested Their model on three datasets of liver, diabetes, and heart. By reducing the number of features, they were able to get good enough accuracies for all three datasets by applying their
machine learning model. The Authors got the best accuracy (78.6%). They were only using structured data, so this will not work for unstructured medical data like patients' interrogation by the doctor, medical reports, etc. Table 3 shows the results and accuracy over the datasets and the reduction in the size of the vector.

Shahadat Uddin [et.al] in 2019 worked on comparative performances of different supervised machine learning algorithms in disease prediction. Regarding the performance of different supervised learning algorithms, DT shows superior results at most times. SVM has been found the least time to show the superior result although it showed the superior accuracy at most times for heart disease, diabetes, and Parkinson's disease. It can be concluded that a single machine learning algorithm may not be the best for all diseases in general, but depending on the disease and datasets, specific ML algorithms perform better as compared to other algorithms.

Jyoti Soni [et.al] have done this research in 2011 to provide a survey of current techniques of knowledge discovery in databases using data mining techniques that are in use in today's medical research particularly in Heart Disease Prediction. A number of experiments have been conducted to compare the performance of predictive data mining technique on the same dataset and the outcome reveals that Decision Tree outperforms and sometimes Bayesian classification is having similar accuracy as of decision tree but other predictive methods like KNN, Neural Networks, Classification based on clustering is not performing well.

K.M. Al-Aidaroos [et.al] have conducted research for the best medical diagnosis mining technique in 2018. For this authors compared Naïve Bayes with five other classifiers i.e. Logistic Regression (LR), KStar (K*), Decision Tree (DT), Neural Network (NN), and a simple rule-based algorithm (ZeroR). For this, 15 real-world medical problems from the UCI machine learning repository (Asuncion and Newman, 2007) were selected for evaluating the performance of all algorithms.

In the experiment, it was found that NB outperforms the other algorithms in 8 out of 15 data sets so it was concluded that the predictive accuracy results in Naïve Bayes is better than other techniques.

Medical Problems	NB	LR	K*	DT	NN	ZeroR
Breast Cancer	72.7	67.77	73.73	74.28	66.95	70.3
Dermatology	97.43	96.89	94.51	94.1	96.45	30.6
Liver Disorder	54.89	68.72	66.82	65.84	68.73	57.98
Haeberman	75.36	74.41	73.73	72.16	70.32	73.53
Hepatitis	83.81	83.89	80.17	79.22	80.78	78.38
Lung Cancer	53.25	47.25	41.67	40.83	44.08	40
Primary Tumor	49.71	41.62	38.02	41.39	40.38	24.78

Table 3: Accuracy of several models on different datasets

III. METHODOLOGY

From an open-source dataset, an excel sheet was created where we listed down all the symptoms for the respective diseases. After which depending on the symptoms, diseases were specified as a part of the dataset. We listed down around 23 diseases with more than 130 unique symptoms in all. The symptoms of an individual were used as input to various machine learning algorithms.

For feature selection, we first selected and removed the symptoms that were either not present in any disease, or were present in all diseases.

After this process, We used the Logistic regression LASSO for model fitting and feature selection. This process removes the features that have the least correlation with the result. After feature selection, our features were reduced from 128 to 70.

As Seen from the literature review, Random Forest and Support Vector machine showed the highest accuracy in cases of normal disease predictions, so, for Classification, we used:

1) *RF:* Random Forest(RF) is an ML algorithm that belongs to the supervised learning technique. It can be used for both Classification problems as well as Regression problems in ML. Random forest takes less training time as compared to the other algorithms. It predicts output with high accuracy and even for the large dataset it can run efficiently. This method is known to maintain accuracy even when a large proportion of data is missing. Random Forest basically works in two-phases: 1st is to create the random forest by combining N decision trees, and 2nd is to make predictions for each tree created in the first phase.
2) *SVM*: Support vector machine algorithm can classify both linear and non-linear data. SVM works by first mapping each data item into an n-dimensional feature space; n being the number of features. Then, The SVM identifies the hyperplane that separates the data items into two classes while maximizing the marginal distance for both classes and minimizing the classification errors. The marginal distance for a class can be defined as the distance between the decision hyper plane and its nearest instance which is a member of that class. To perform the classification, we then need to find the hyper plane that differentiates the two classes with maximum margin.

IV. RESULTS

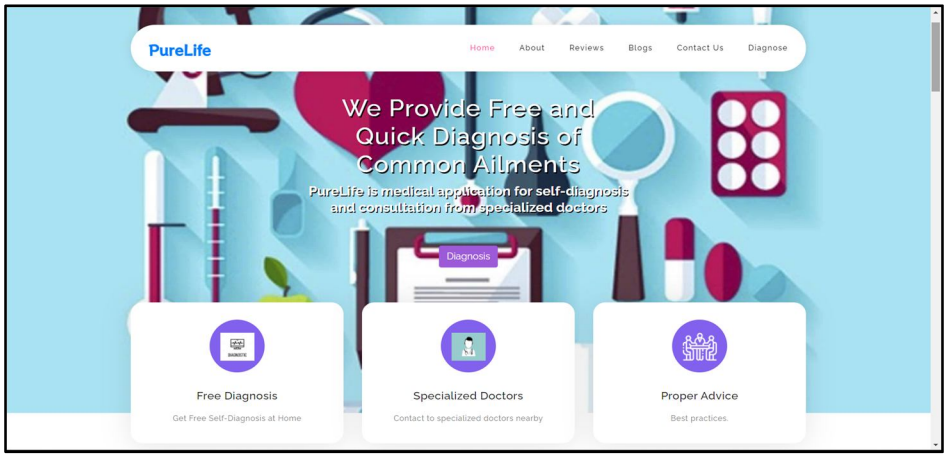


Fig 1: Landing page of our webApp PureLife

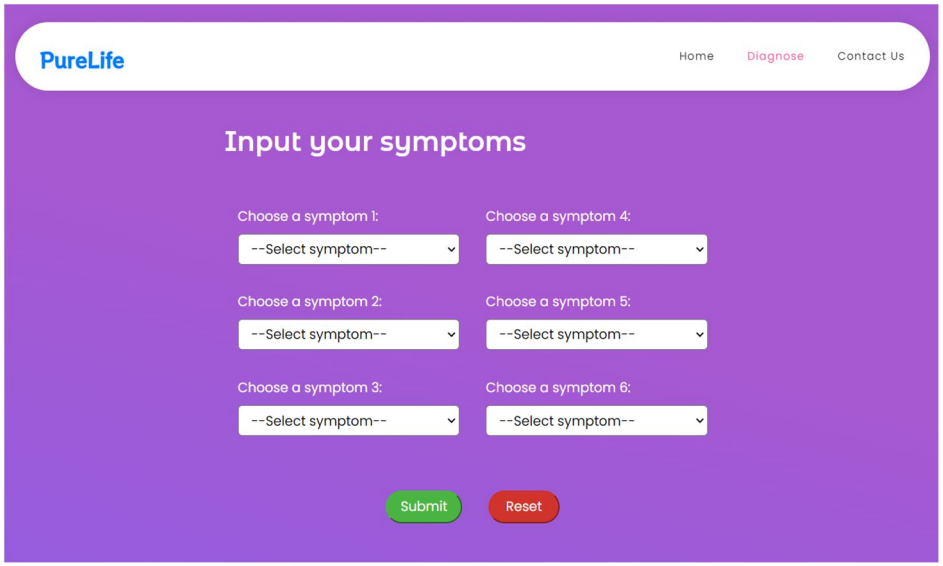


Fig 2: Input page for symptoms

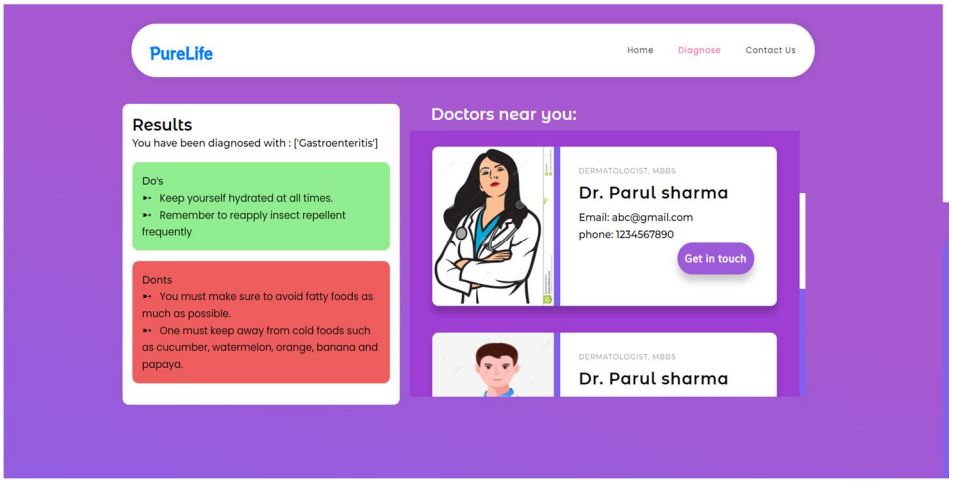


Fig 3: Get specialized doctors & information regarding predicted disease.

Feature selection offers a simple yet effective way to eliminate redundant and irrelevant data. Removing the irrelevant data improves learning accuracy, reduces the computation time, and facilitates an enhanced understanding for the learning model or data. We applied feature selection at two stages and reduced the features from 128 to 70. After processing the data, we passed them into our machine learning models and the results are as followed:

Algo-rithm Used	Advantages	Limitations	Accuracy
RF	It is basically a collection of Decision trees and hence it's proven to be better than that, while staying easy to grasp.	More complex and computationally expensive.	84%
		Number of base classifiers need to be defined.	
	Scales easily and efficiently for big datasets	Greater chances of overfitting over a given dataset.	
	It can provide estimates of what variables or attributes are important in the classification	It favors those variables or attributes that can take a high number of different values in estimating variable importance.	
SVM	More robust as compared to Logistic Regression.	Computationally expensive for large and complex datasets	90%
	Can handle multiple feature spaces.	May underperform in cases where the dataset has noise.	
	Lesser the chances of overfitting over a dataset.	The model created as well as the weight of the variables can be difficult to comprehend	
	Comparatively efficient for classifying semi-structured or unstructured data such as images and texts.	Generally, an SVM model cannot classify more than two classes unless extended.	

Table 4: Advantages, LImitations and accuracy of the models used

V. DISCUSSION

The field of study for medical science as well as machine learning is very vast, with continuous developments. We have only worked on and deployed SVM and RF models. In the future, we will study and implement other machine learning techniques. The accuracy of the model is very crucial in the medical field, and so, we aim to improve the accuracy of the model even further, by implementing different algorithms and improvising our dataset.

Our project recommends specialized doctors that are near the user. We aim to provide a platform on which the user can connect with the doctors as well as pharmacy stores.

We aim to deliver medical professionalism with ease of access and accuracy to every citizen.

VI. LIMITATIONS

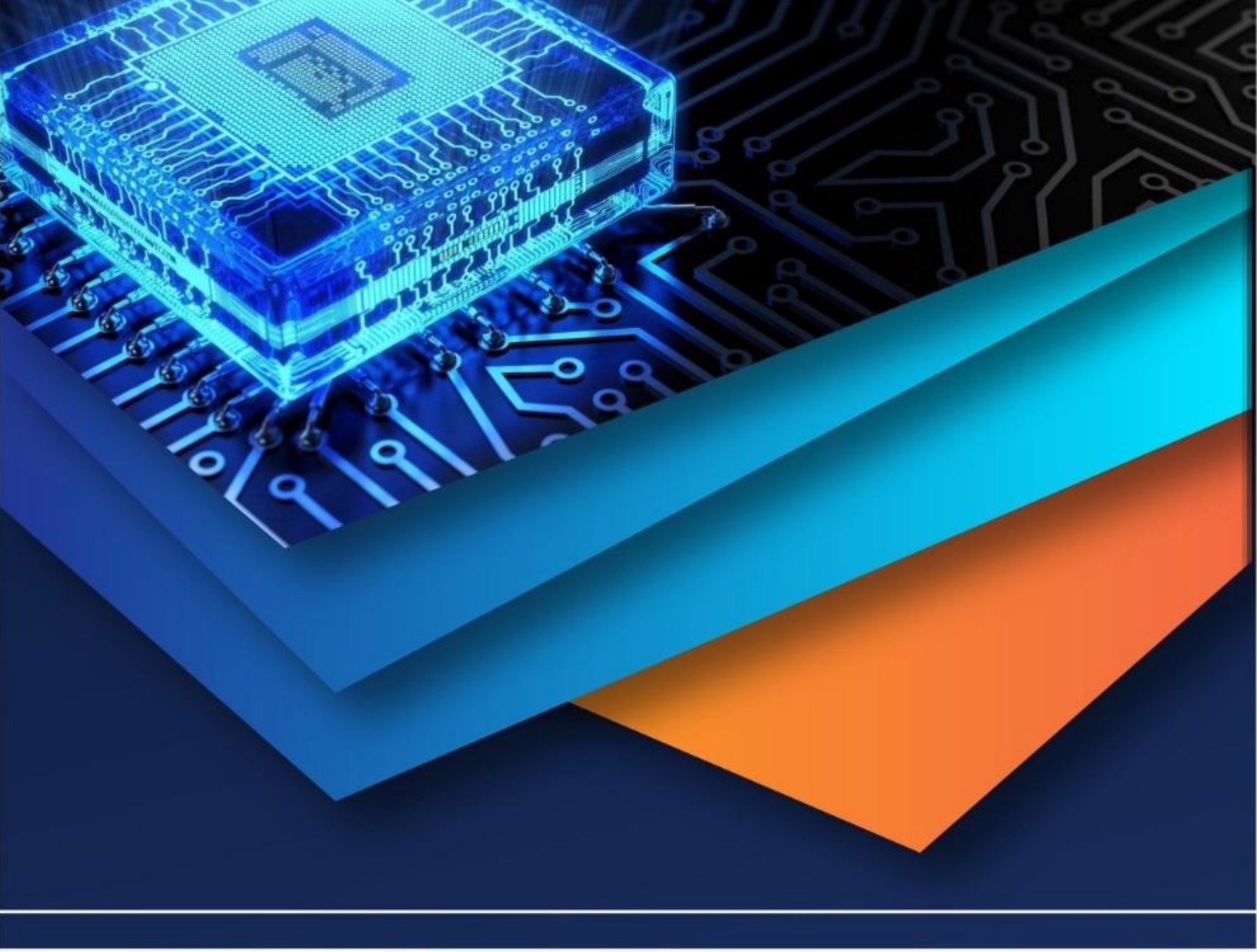
We were not able to gather data and accuracy of models regarding common ailments specifically.
We cannot add real time diseases for which the symptoms keep evolving, such as the covid-19.
Our prediction system can prove to be helpful and can be used in the diagnosis of a disease in case of an emergency, and for regions where sufficient facilities and resources are unavailable, However, doctors and medical professionals are always recommended to diagnose and treat the ailments.

VII. CONCLUSION

The research paper presented the technique of predicting the disease based on the symptoms of an individual patient. We used different machine learning models on the same dataset to compare the results and provide the best solution to the problem. The SVM model gave the highest accuracy of 90 % for the prediction of diseases using the above-mentioned factors. Other ML algorithms also gave good accuracy values but as SVM performed, we used its prediction. Once the disease is predicted, we could easily manage the medications and treatments. Aside from its main functionality, the future scope in this project is vast and can include means of a better platform to connect and provide for the patients with specialized recommended doctors, as well as connecting the pharmacy stores to provide for the medicines as well.

REFERENCES

[1] Iwendi C, Bashir AK, Peshkar A, Sujatha R, Chatterjee JM, Pasupuleti S, Mishra R, Pillai S and Jo O, “COVID-19 Patient Health Prediction Using Boosted Random Forest Algorithm”. (2020) Front. Public Health 8:357. DOI: 10.3389/fpubh.2020.00357
[2] Keniya Rinkal and Aman Khakharia and Vruddhi Shah and Vrushabh Gada and Ruchi Manjalkar and Tirth Thaker and Mahesh Warang and Ninad Mehendale “Disease Prediction from Various Symptoms Using Machine Learning”. (2020) SSRN: 3661426.
[3] Anant Agrawal, Harshit Agrawal, Shivam Mittal, Mradula Sharma, “Disease Prediction Using Machine Learning”. 3rd International Conference on Internet of Things and Connected Technologies (ICIoTCT), ISSN: 1556-5068. 2018.
[4] Shahadat Uddin, Arif Khan, Md Ekramul Hossain, and Mohammad Ali Moni, “Comparing different supervised machine learning algorithms for disease prediction”, BMC Medical Informatics and Decision Making (2019) 19:281.
[5] Jyoti Soni, Ujma Ansari, Dipesh Sharma, and Sunita Soni, “Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction”. (2011) International Journal of Computer Applications. DOI: 10.5120/2237-2860.
[6] K.M. Al-Aidaroos, A.A. Bakar, and Z. Othman, “Medical Data Classification With Naive Bayes Approach”. (2012) International Technology Journal 11. DOI: 10.3923/itj.2012.1166.1174.
[7] Sayantan Saha, Argha Roy Chowdhuri et al., "Web Based Disease Detection System", IJERT, vol. 2, no. 4, April 2013, ISSN 2278-0181.
[8] Shadab Adam et al., "Prediction system for Heart Disease using Naïve Bayes", International Journal of Advanced Computer and Mathematical Sciences, vol. 3, no. 3, pp. 290-294, ISSN 2230-9624 (2012).
[9] S. Jadhav, R. Kasar, N. Lade, M. Patil, and S. Kolte, “Disease Prediction by Machine Learning from Healthcare Communities,” InternationalJournal of Scientific Research in Science and Technology, pp. 29–35,2019
[10] Dhenakaran, K. Rajalakshmi Dr SS. "Analysis of Data Mining Prediction Techniques in Healthcare Management Systems." International Journal of Advanced Research in Computer Science and Software Engineering (2015).
[11] N. Skyttberg, J. Vicente, R. Chen, H. Blomqvist, and S. Koch, "How to improve vital sign data quality for use in clinical decision support systems? A qualitative study in nine Swedish emergency departments," BMC medical informatics and decision making, vol. 16, p. 1, 2016
[12] A. Wright, T.-T. T. Hickman, D. McEvoy, S. Aaron, A. Ai, J. M. Andersen, et al., "Analysis of clinical decision support system malfunctions: a case series and survey," Journal of the American Medical Informatics Association, p. ocw005, 2016.
[13] C. Y. Tsai, S.H. Wang, M.H. Hsu, and Y.C. J. Li, "Do false positive alerts in naïve clinical decision support systems lead to false adoption by physicians? A randomized controlled trial," Computer Methods and Programs in Biomedicine, vol. 132, pp. 83-91, 2016.
[14] M. El-Bardini and A. M. El-Nagar, "Direct adaptive interval type-2 fuzzy logic controller for the multivariable anesthesia system," Ain Shams Engineering Journal, 2011.
[15] L. Qiao and G. D. Clifford, "Suppress False Arrhythmia Alarms of ICU Monitors Using Heart Rate Estimation Based on Combined Arterial Blood Pressure and Ecg Analysis," in Bioinformatics and Biomedical Engineering, 2008. ICBBE 2008. The 2nd International Conference on, 2008, pp. 2185-2187.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY