https://doi.org/10.48047/AFJBS.6.14.2024.5059-5069

AFJBS

African Journal of Biological Sciences

Journal homepage: http://www.afjbs.com

Research Paper

Open Access

Smart Health Care System Using Data Mining and Visualization Techniques

Suraj Kumar Tellakula[1], Vachaspathi Gnaneswar Garlapati[2], Venkataramana Baratam[3], Rohith Venkata Sai Kunta[4], Dr. Vamsidhar Enireddy[5]

[1]Department of Computer Science and Engineering, Koneru Lakshmaiah Educational Foundation, Guntur, India. surajkumar.ayanokoji.7@gmail.com
[2]Department of Computer Science and Engineering, Koneru Lakshmaiah Educational Foundation, Guntur, India. gvachaspathignaneswar@gmail.com
[3]Department of Computer Science and Engineering, Koneru Lakshmaiah Educational Foundation, Guntur, India. bvenkataramana2852@gmail.com
[4]Department of Computer Science and Engineering, Koneru Lakshmaiah Educational Foundation, Guntur, India. kvenkatasai321@gmail.com
[5]Department of Computer Science and Engineering, Koneru Lakshmaiah Educational Foundation, Guntur, India. enireddy.vamsidhar@gmail.com

Volume 6, Issue 14, Aug 2024

Received: 15 June 2024

Accepted: 25 July 2024

Published: 15 Aug 2024

doi: 10.48047/AFJBS.6.14.2024.5059-5069

Abstract-Technologies such as data mining and data visualization improve healthcare in the modern world. Hospital wait times can be lengthy, and after a prescription is filled, there may be additional costs. This can lead to costly hospital expenses. The goal of this research is to develop a health prediction system that gathers medical information and symptoms, then uses that information to anticipate the disease. In addition, it provides disease prognoses, medical advice, and a platform that matches patients with qualified physicians based on proximity and experience. In order to analyze and find patterns and insights in large patient datasets, the study focuses on understanding data mining techniques, such as the K-Nearest Neighbor Method, Decision Tree Algorithm, Random Forest Classifier, Logistic Regression & Support Vector Machine Algorithm, as well as data visualization techniques. Algorithms are employed to help in diagnosing different illnesses such as heart disease, kidney disease, diabetes, and breast cancer. After preprocessing the datasets, the aforementioned algorithms are applied to various diseases, and evaluation metrics are calculated for each algorithm and disease. Based on the results, accuracy was measured to find the robust algorithm suitable for diagnosing most of the diseases. The findings revealed that the algorithm Random Forest achieved the highest accuracy among all algorithms for each of the diseases mentioned and it is clear from the results that the algorithm has demonstrated its reliability when compared to other algorithms, this further strengthens the effectiveness of the algorithm in diagnosing diseases within the Smart Healthcare System.

Keywords: Medical Research Enhancement, Data Mining, Visualization Approaches, Early Disease Diagnosis, Personalised Healthcare, Predictive Analytics in Healthcare, and Smart Healthcare Systems.

1. Introduction

Millions of gigabytes of data are collected and stored in databases worldwide, with this volume is expected to grow [1]. Over the past year, there has been a noticeable increase in patients in the healthcare industry. Delivering top-notch healthcare requires making prompt treatment decisions and accurate diagnosis. A web-based system or algorithms can provide a strong economy of scale while optimizing the delivery of healthcare.

A number of data mining techniques are incorporated in order to create a successful smart health prediction system, such as the Random Forest Classifier, Logistic Regression, K-Nearest Neighbor Method, Decision Tree Classifier, & Support Vector Machine method. The forementioned algorithms are designed for healthcare to handle patient records, doctor information, and disease prognosis. In the research, there's a particular focus on supervised learning using prepared data, aiming to attain trustworthy outcomes by utilizing optimal datasets.

A powerful tool for swiftly comprehending complex data sets and gaining insightful knowledge is visualization. Methods like tree maps, mosaics, and graph-based, hierarchical techniques like scatter plot matrices are employed. Researchers and medical practitioners can discover relationships in medical datasets using these visualization techniques, which aids in improving clinical judgment and cutting costs. Charts and graphs make it simple for healthcare professionals to obtain critical information.

Identifying and foreseeing diseases becomes simpler using our platform. Advanced data mining methods and data visualization approaches are combined in this platform. By identifying health issues early on, the user-friendly system will reduce the number of needless hospital visits and associated expenses. Additionally, individualized dietary and medication advice will be provided. Visualization of data and predictive analysis allow users to manage their health actively, optimizing resources and reducing the cost of needless doctor visits.

2. Literature Survey

In a presentation, Basma Boukenze, Hajar Mousannif, and Abdelkrim Haqiq examined the use of data mining from sources such as electronic medical records to investigate predictive analytics in the healthcare industry. Additionally, by using the C4.5 decision tree classifier on the Weka platform to predict chronic kidney illnesses with high precision, sensitivity, and accuracy, they have highlighted the importance of big data analytics and machine learning. In conclusion, it makes the case that machine learning ought to advance in order to acquire knowledge that will allow intelligence to address health-related challenges [4].

The suggested paper by N. Shabaz Ali and G. Divya examines data mining for illness prediction based on user symptoms, highlighting the importance of early detection in saving lives. Important algorithms including Naive Bayes, decision trees, Random Forest, Nearest Neighbor, Neural Networks, Support Vector Machines, Logistic/Linear Regressions, and Discrimination Analysis are listed. The accurate prediction percentages for specific diseases are underlined, indicating the common use of neural network methods and support vector machines in healthcare anticipation [5].

"Developing Smart E-health Prediction Systems" by K. Bala Sita Ramireddy, K. Vasanthi, G. Pooja Reddy, Ravi Kumar Tenali, and M. Trinath Basu was specifically looking through the data mining techniques to see how clinical predictions can be improved. The system christened as Smart Health Prediction System provides real-time health advice via an online interface that is easy to use for all people. It utilizes decision tree algorithms, particularly the Naïve Bayes classifier, and is developed using Java and Eclipse IDEs. The paper ends by noting the system's potential future applications and outlining the benefits of using data mining in the medical industry [6].

The presentation by Amanze, Bethran Chibuike, and Oguji Francis Chikezie emphasized the value of data analytics and visualization in the medical field and offered a practical illustration of how these tools might be applied to the analysis of medical data. It also examines previous research on web-based visualization systems and related studies that have used knowledge management and data mining to solve issues. Additionally, a Python software is used in this case study to do data analytics and visualization on a dataset related to heart illness, resulting in graphical reports that include charts and graphs. Additionally mentioned is the use of SPSS for extracting data from datasets [7].

The advantages of data mining in healthcare are introduced in the paper by T Kavya and Dr. K Santhi Sree, with a focus on fraud detection, economy of scale decisions, and tailored health profiles. To provide instantaneous online therapy recommendations using decision trees or Naive Bayes algorithms for individuals unable to visit hospitals. The system schedules consultations with specialists and recommends medications when necessary. The paper ends by highlighting the potential advantages of data mining for medical purposes while taking security and privacy into account [8].

3. Methodology

In the past, receiving effective treatment was challenging and took a long time because doctors were unable to correctly identify the illness's underlying cause and provided insufficient support. Better sickness treatment is easier to achieve with this smart healthcare system, which makes use of data mining and data visualization tools. The algorithms used offer predictive outcomes to aid doctors in diagnosing and treating patients. The workflow diagram that follows demonstrates how this healthcare system functions, making it more user-friendly and effective.

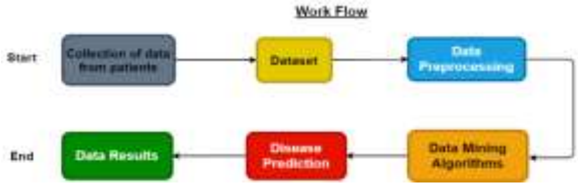


Figure 1. Workflow of a Smart Healthcare System

Workflow Steps:

1. **Collection of Data from Patients**: The process begins by gathering relevant data from patients.
2. **Dataset Creation**: The data that has been **gathered** is sorted and arranged into a dataset.
3. **Data Preprocessing:** This step involves **cleaning**, transforming, and preparing the dataset to make it ready for analysis.
4. **Application of Data Mining Algorithms:** The pre-processed data is subjected to a variety of data mining methods, including Random Forest method, KNN algorithm, SVM technique, Decision Tree method, and Logistic Regression.
5. **Disease Prediction:** The algorithms predict diseases based on the input features.
6. **Data Results:** The outcomes of disease prediction are obtained.

Logistic Regression

Logistic regression is used in Smart Healthcare Systems to forecast the existence or non-existence of a disease (independent variable) based on test results and symptoms (dependent variable). Initially, the dataset undergoes preprocessing to tidy up the data and manage any missing values. Then, the dataset is divided into training and testing sets using the sklearn library. The training data is utilized to teach the logistic regression model, which focuses on binary classification, distinguishing outcomes as either 0 or 1. Finally, the model's effectiveness is assessed using the testing data. The logistic regression formula is expressed as:

$$f(k) = \frac{1}{1+e^{-k}} = \frac{1}{1+e^{-(b_1.x_1+\cdots+b_n.x_n+a)}} \qquad (1)$$

Where, $k = b_1.x_1+b_2.x_2+\ldots\ldots+b_n.x_n+a$

$e = 2.71828$

$f(k)$ = output of logistic regression

b = weights

x = features

a = intercept term

After being trained, the model examines fresh data and forecasts the likelihood of illness occurrence depending on input factors.

Decision Tree Induction

Begin by examining patient data and choosing important characteristics like age, gender, symptoms, and certain medical measurements. Construct a tree layout using an algorithm where each step represents a choice based on a particular characteristic. To arrive at a final forecast and create predictions for new patient data, the tree is used. Based on the patient's data, the trained model can forecast the probability that the patient will have a disease. However, it's worth noting that decision trees have constraints, like needing discrete target attributes for approaches like ID3 and C4.5.

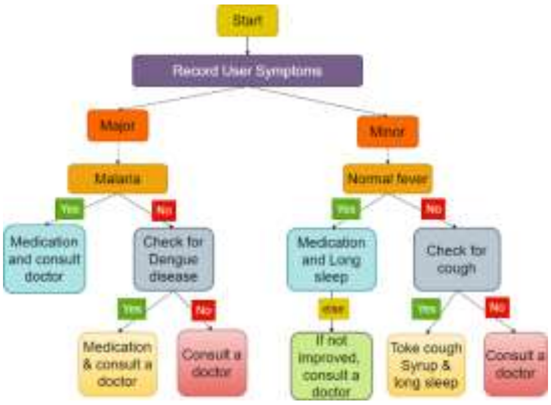


Figure 2. Example on Decision Tree

The above decision tree describes if the symptoms indicate a major issues like malaria or dengue, it advises medication and seeing a doctor. For minor concerns like normal fever or cough, it suggests rest and medicine, with a doctor's visit if needed. Consulting a doctor is crucial for accurate diagnosis and safe treatment.

Support Vector Machine

Data is pre-processed with Support Vector Machine to remove missing values and clean up dirty data. Next the data is divided into categories based on the classification, with ‘negative’ denoting the "absence of disease" and ‘positive’ denoting the "presence of disease". During training, the Support Vector Machine algorithm finds the best line or plane that divides the data, and support vectors are the data points closest to that line or plane. Two methods can be utilized to implement Support Vector Machines: mathematical programming and kernel functions.

The SVM equation is formulated as:

$$f(x) = sign(b + w^T) \qquad (2)$$

Where, f(x) = The predicted category or class label.

x = input feature vector

b = bias term

w = weight vector

sign(.) = sign function (+1, -1)

Once the model is trained, it is assessed on fresh data to measure how well it performs, to determine how well it predicts diseases metrics such as recall, accuracy, precision, and F1 score are incorporated.

K-Nearest Neighbor

In terms of regression and classification, K-Nearest Neighbor (KNN) is one of the most popular algorithms relying on a supervised approach. The first step involves gathering and preparing the dataset, which includes tasks like cleaning and dealing with missing data. The parameter 'K', which indicates the number of nearest neighbors, holds significance and is usually selected in advance. KNN algorithm tends to memorize the entire dataset instead of generalizing well during the training phase. It calculates the distance between each new data point and the training points, finds the K closest neighbors, makes predictions, and categorizes the neighbors based on the most common type. The general formula for the KNN is:

$$\boldsymbol{d = \sqrt{\sum_{i=1}^{n} \frac{(x_i - y_i)^2}{R_i^2}}} \qquad \mathbf{(3)}$$

Where in KNN Algorithm,

d = distance between the two data points

R_i = i^{th} feature's scaling factor.

n = number of characteristics

$x_i = y_i = i^{th}$ feature values for the two data points.

When the decision lines are not straight, K-Nearest Neighbor can be helpful for predicting diseases, but it might struggle with datasets that have a lot of dimensions.

Random Forest

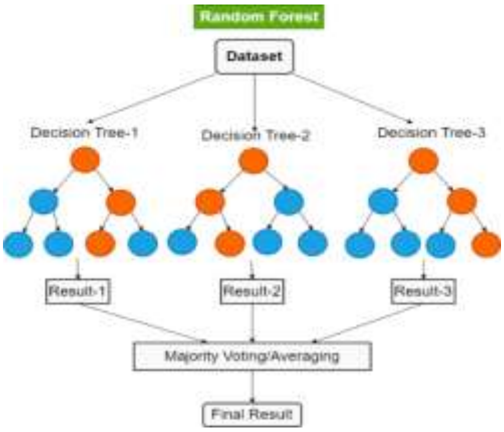


Figure 3. Example on Random Forest

To arrive at a single result, the output of several decision trees is combined to create a Random Forest. Before training, the technique needs to define the number of trees, node size, and number of features sampled. It builds decision trees using bootstrapped sampling and reduces overfitting and bias.

This image depicts how a dataset generates multiple decision trees that process the data independently to produce results. To produce a final result, majority voting or averaging is used to combine the individual results and also this diagram illustrates how multiple decision trees can collaborate to make better predictions. Random Forest deals with missing information and is assessed using measures like accuracy.

4. Results & Outputs

Data mining algorithms were used in this study to forecast conditions like diabetes, kidney disease, heart disease, and breast cancer. For every ailment, a different method was used, such as the Random Forest Classifier, K-Nearest Neighbor Method, Decision Tree Method, Support Vector Machine Algorithm, and Logistic Regression Technique. Their predictive performance was assessed by examining precision, recall, F1-Scores, accuracy, and support parameters. F1-Score metric evaluates the balance between recall metrics and precision. Recall demonstrates the capacity to recognize important observations, whereas Precision displays the accuracy of positive predictions. While Accuracy assesses the overall accuracy of predictions, Support metrics quantify the number of occurrences in each class. The performance of each algorithm for a range of diseases is displayed in the table below.

Certain data mining models may be specifically implied for particular diseases, posing challenges when attempting to apply them to diverse health conditions. Consequently, a range of algorithms were employed to assess their performance across various diseases. Presented below are the graphical representations of different algorithms' outcomes when applied to the following health issues:

1. Cardiac conditions
2. Diabetes mellitus
3. Breast carcinoma
4. Renal ailments.

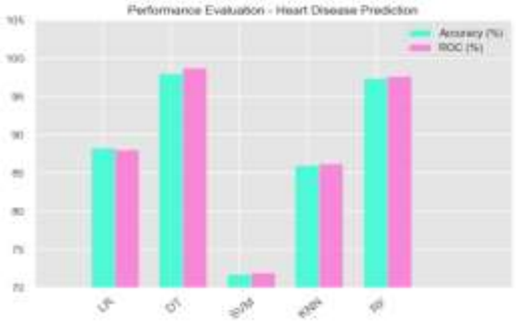


Figure 4. Heart Disease Prediction

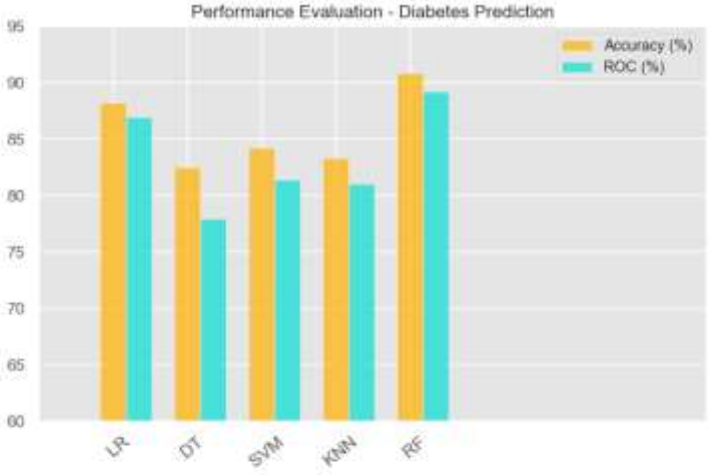


Figure 5. Diabetes Prediction

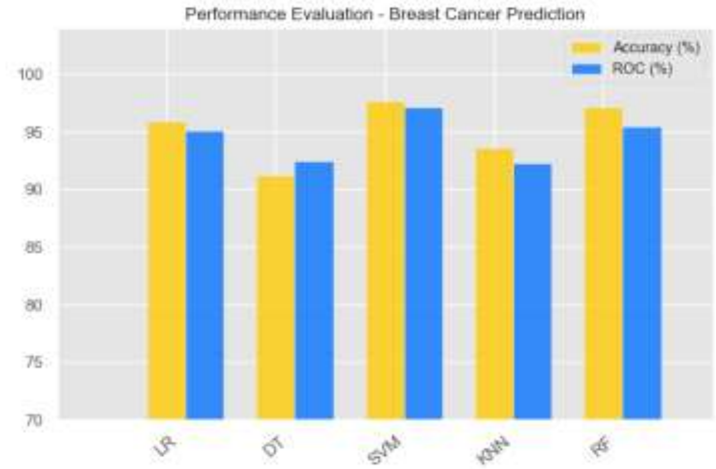


Figure 6. Breast Cancer Prediction

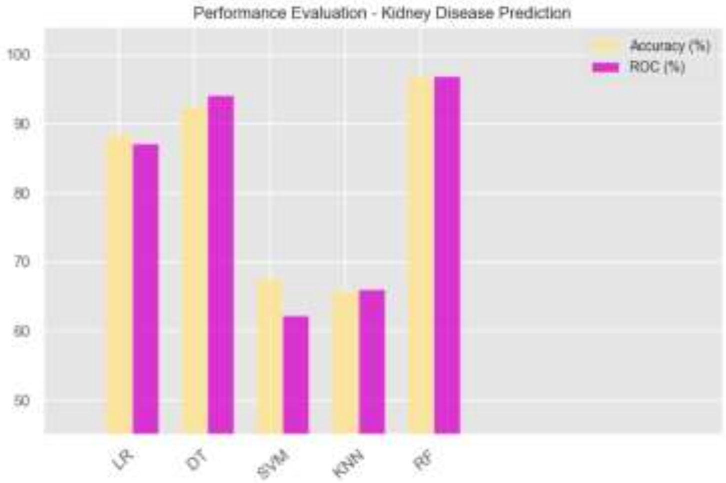
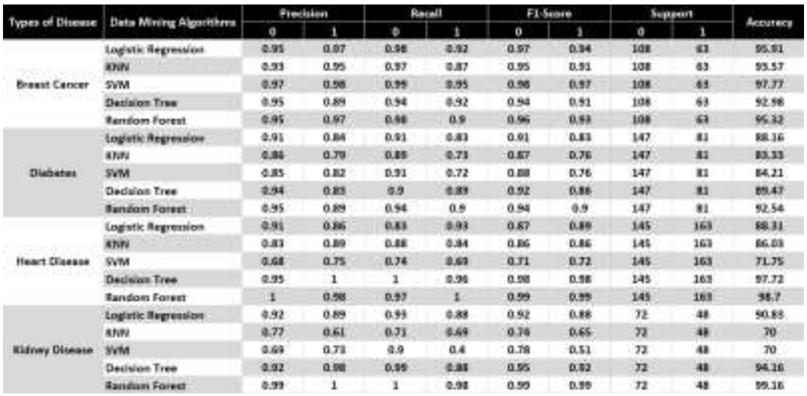


Figure 7. Kidney Disease Prediction

Furthermore, there is a table that shows the evaluation metrics of various algorithms for these illnesses.

5. Conclusion



Types of Disease	Data Mining Algorithms	Precision		Recall		F1-Score		Support		Accuracy
		0	1	0	1	0	1	0	1	
Breast Cancer	Logistic Regression	0.95	0.07	0.98	0.92	0.97	0.94	108	63	95.91
	KNN	0.93	0.95	0.97	0.87	0.95	0.91	108	63	93.57
	SVM	0.97	0.98	0.99	0.95	0.98	0.97	108	63	97.77
	Decision Tree	0.95	0.89	0.94	0.92	0.94	0.91	108	63	92.98
	Random Forest	0.95	0.97	0.98	0.9	0.96	0.93	108	63	95.32
Diabetes	Logistic Regression	0.91	0.84	0.91	0.83	0.91	0.83	147	81	88.16
	KNN	0.86	0.79	0.89	0.73	0.87	0.76	147	81	83.33
	SVM	0.85	0.82	0.91	0.72	0.88	0.76	147	81	84.21
	Decision Tree	0.94	0.83	0.9	0.89	0.92	0.86	147	81	89.47
	Random Forest	0.95	0.89	0.94	0.9	0.94	0.9	147	81	92.54
Heart Disease	Logistic Regression	0.91	0.86	0.83	0.93	0.87	0.89	145	163	88.31
	KNN	0.83	0.89	0.88	0.84	0.86	0.86	145	163	86.03
	SVM	0.68	0.75	0.74	0.69	0.71	0.72	145	163	71.75
	Decision Tree	0.95	1	1	0.96	0.98	0.98	145	163	97.72
	Random Forest	1	0.98	0.97	1	0.99	0.99	145	163	98.7
Kidney Disease	Logistic Regression	0.92	0.89	0.93	0.88	0.92	0.88	72	48	90.83
	KNN	0.77	0.61	0.71	0.69	0.74	0.65	72	48	70
	SVM	0.69	0.73	0.9	0.4	0.78	0.51	72	48	70
	Decision Tree	0.92	0.98	0.99	0.88	0.95	0.92	72	48	94.16
	Random Forest	0.99	1	1	0.98	0.99	0.99	72	48	99.16

Figure 8. Classification Table

Data mining algorithms have shown some encouraging outcomes in this research towards developing a smart healthcare system. A variety of algorithms have been used such as Support Vector Machine Algorithm, Decision Tree Algorithm, Random Forest Classifier, K- Nearest Neighbor Method, and Logistic Regression Technique to forecast the illnesses like diabetes, kidney disease, heart illness, and breast cancer.

According to the evaluation criteria, Random Forest works well for diabetes prediction, whereas Support Vector Machine is a good choice for breast cancer prediction. Cardiovascular illness can be effectively managed with the use of Random Forest algorithms and Decision Trees. In particular, Random Forest performs exceptionally well in kidney illness prediction. Every illness requires a unique course of treatment. The incorporation of these algorithms into the healthcare system requires ongoing testing and optimisation.

As a result of these results, Random Forest seems to be the most accurate algorithm across a wide range of disease types in the Smart Healthcare System. The decision tree algorithm also performed well for three out of the four diseases. SVM algorithms, KNN algorithms, and logistic regression algorithms performed less well than Random Forest techniques and Decision Tree algorithms.

6. Future Scope

In the future, researchers will concentrate on finding and picking out important information from deep learning methods to improve the efficacy and to make the diagnostic results better and more accurate in smart healthcare systems. Moreover, it is important to consider the additional factors for finding accuracy when selecting data mining algorithms for real-world applications specifically the significance of computational efficiency, scalability, Cross-validation, and interpretability. As part of the future work, exploring these aspects further and implementing them into the research will increase the practical applicability of the findings. Aiming to explore more algorithms in data mining to enhance the performance of the Smart Healthcare System. Using more advanced techniques for digging through data can make diagnostic outcomes better when there's a wide variety of information in the training set.

References

[1] Ekwonwune, E. N., Ubochi, C. I., & Duroha, A. E. (2022). Data mining as a technique for healthcare approach. International Journal of Communications, Network and System Sciences, 15(09), 149-165.

[2] Kunjir, A., Sawant, H., & F., N. (2016). A review on prediction of multiple diseases and performance analysis using data mining and visualization techniques. International Journal of Computer Applications, 155(1), 34-38.

[3] Wibamanto, W., Das, D., & Chelliah, S. A. (2020). Smart health prediction system with data mining. *International Journal of Current Research and Review, 12*(23), 14-19.

[4] Boukenze, B., Mousannif, H., & Haqiq, A. (2016). Predictive analytics in healthcare system using data mining techniques. Computer Science & Information Technology (CS & IT).

[5] Shabaz Ali, N., & Divya, G. (2020). Prediction of Diseases in Smart Health Care System using Machine Learning. International Journal of Recent Technology and Engineering, 8(5), 2277-3878.

[6] Pooja Reddy, G., Trinath Basu, M., Vasanthi, K., Bala Sita Ramireddy, K., & Ravi Kumar Tenali (2019). Smart E-Health Prediction System Using Data Mining. International Journal of Innovative Technology & Exploring Engineering, 8(2), 2278-3075.

[7] Amanze Bethran Chibuike & Oguji Francis Chikezie (2022). Data Analytics and Visualization in the Health Sector. International Journal of Trend in Research and Development.

[8] Kavya, T., & Santhi Sree, K., (2021). Smart Health Prediction Using Data Mining. Journal of Emerging Technologies and Innovative Research, 8(6), 892-899.

[9] Nikita Kamble, Manjiri Harmalkar, Manali Bhoir, & Supriya Chaudha (2017). Smart Health Prediction System Using Data Mining. International Journal of Scientific Research in Computer Science, Engineering, and Info.

[10] Ahmad, P., Qamar, S., & Qasim Afser Rizvi, S. (2015). Techniques of data mining in healthcare: A review. International Journal of Computer Applications, 120(15), 38-50. Information Technology, 02(02), 1020-1025.