

## Disease Prediction Based on User Symptoms Using Machine Learning Algorithms

<sup>1</sup>G. Geetha Devi, <sup>2</sup>Palepu Prasanna, <sup>3</sup>Pullouri Sravanthi, <sup>4</sup>Gandla Sharanya

<sup>1</sup> Assistant Professor, Department of Information Technology, Bhoj Reddy Engineering College for Women, Vinay Nagar, Hyderabad

<sup>2,3,4</sup> Student, Department of Information Technology, Bhoj Reddy Engineering College for Women, Vinay Nagar, Hyderabad

**Abstract**—Every day, many individuals encounter different illnesses. The prognosis of a disease is the most pivotal part of treatment. Enormous increase in healthcare and medical data enabled accurate medical data analysis, which aids in early sickness discovery and beforehand patient care. This study focuses on performing research on the enormous medical data by exercising numerous supervised classification algorithms like Decision Tree, Support Vector Machine, K- Nearest Neighbor, Logistic Regression, Naive Bayes and Random Forest to anticipate the most probable disease grounded on the symptoms and also to directly prognosticate the possibility of whether or not the person might be suffering from that particular illness. Based on the symptoms, the model uses the results of the supervised classification algorithms and gives a final validation indicating the disease that the existent might be suffering. On combining the prognostications from all the below classification algorithms, the model returns a more accurate verification when compared to the prognostications made by individual models. This study enhances the swiftness of decision-making and can reduce the rate of false cons. It helps healthcare associations make better decisions about how to proceed with early patient care. It also aids healthcare professionals in developing further effective ways of treating patients.

**Keywords**—Decision Tree, Support Vector Machine, K-Nearest Neighbor, Logistic Regression, Random Forest, Naïve Bayes

### I. INTRODUCTION

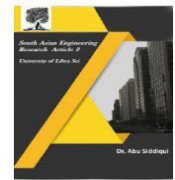
Around 55 million deaths are recorded every year globally, among which 55% of the deaths were caused due to various diseases suffered by individuals. In addition, 7 out of 10 deaths are caused by non-communicable conditions, which account for 74% of deaths as per the data records. Many people experience different kinds of diseases now and then. With the emergence of the digital age and technological advancements, a colossal amount of multifaceted patient data is developed, including clinical factors, diagnostic information, patients' records, medical research, and paraphernalia. Prediction of a disease is the most crucial part of treatment. The most challenging task that doctors often deal with is struggling to make accurate diagnoses based entirely on the patient's symptoms [11]. Hence, Machine learning plays a significant role in forecasting, to complete this important and demanding task of predicting illness. The

model utilizes medical data to identify patterns grounded on the symptoms. The users can select the symptoms they are suffering from and obtain a prediction of the most probable disease that they might be suffering from based on the symptoms. It performs a practical medical data analysis and predicts the disease before exacerbating the patient's health, allowing early patient care.

### II. RELATED WORKS

[1] This paper provides early disease prediction and medical benefits using available medical data. This paper uses supervised machine learning algorithms to predict the most probable disease the users could possibly have by taking their symptoms as inputs. The algorithms used were Naïve Bayes, RF, KNN, SVM and Decision Tree. The system is developed such that Naïve Bayes is used for disease prediction, KNN for classification, Logistic Regression for extracting the features and Decision tree for dividing the dataset into smaller parts. Measures like recall, precision and F1 were used to obtain an accuracy of 71.28% for KNN, 84.5% for Decision tree, 98.95% for Random Forest, 89.4% for Naive Bayes and 96.49% for SVM. The authors concluded that the disease predictor was developed using the grails framework and obtained a system accuracy of 98.3%. This system uses the grail framework for disease predictors.

[2] The paper focuses on providing a solution using a web/android application such that a user can access the application remotely when the doctors are unavailable. The application solves minor problems, but users must visit a hospital in person for a more thorough examination. The system focuses on providing users with rapid and accurate illness prediction based on symptoms and the severity of the sickness projected. The system compares the symptoms provided with the information within the dataset. If the symptoms match the dataset, it provides the relevant disease or notifies it as a 'wrong symptom'. Then the prompt would ask if the user would like to save the symptoms within the information. Naive Bayes Algorithm, KNN Algorithm, Decision Tree Algorithm, RF, and SVM were employed in the system to provide accurate predictions. The system is then taught in a web application using python and Django so that the users can get solutions without the hassle of visiting the doctor.



[3] The paper focuses on providing a virtual, precise and early examination of any health-related problems. The conventional ways require a lot of time to visit the doctor and come to a solution. Thus, to make it less time-consuming, the paper focuses on what the individual might be suffering from with the help of the information given in the dataset. The dataset contains about 230 diseases for processing. The system shows the disease as a result of predicting the disease the individual may be suffering from, based on the symptoms, age, and gender of the individual. The dataset was then processed using several ML models like Fine, Medium and Coarse Decision trees; Kernel and Gaussian Naive Bayes, Fine, Medium, Coarse, Weighted and Subspace KNN, and RUS Boosted trees. The accuracy varied for each model, and the model with the highest accuracy was used to build the project. The Weighted KNN algorithm had the highest accuracy with 93.5%, followed by Fine KNN, having an accuracy of 80.3%.

[4] The paper focuses on predicting heart diseases using supervised learning algorithms. The report aims to provide better accuracy to predict the chance of heart diseases based on the features such as Age, Cholesterol, Chest pain and Blood pressure. Algorithms such as Random Forest Classifier, Support Vector Classifier [13], KNN Classifier, Decision Tree, Naive Bayes Classifier and Logistic Regression were implemented, and various data mining techniques were used to classify the patient risk level. KNN classifier has obtained the highest accuracy of 87%.

The algorithms used in [5] were Decision Tree, K-Nearest Neighbour, Random Forest and Naive Bayes to develop a disease prediction system where each algorithm was considered as an individual model and the predictions made by the respective individual models were displayed. These models showed signs of better accuracy for producing an estimated result.

[11] This paper emphasizes on the detection of possible chronic diseases like diabetes, cancer, arthritis and other diseases the user may be suffering from, based on the symptoms provided by him. It was exclusively created for end users' usage only. The authors utilized real hospital data to predict the most accurate disease which is in structured and textual format. The model makes use of machine learning algorithms like Naive Bayes for predicting the disease, KNN algorithm for clustering and Logistic Regression for final output which will be in the form of zeroes and ones. Evaluation methods like accuracy [19], Recall and F1-Measure are used to calculate the performance.

[14] This paper focuses on two types of analysis: structured analysis and unstructured analysis. The author focuses on identifying chronic diseases in a particular region and community using structured analysis. In unstructured analysis, the author used Random Forest, Naïve Bayes Classifiers, Decision tree and K-Nearest Neighbor algorithms to predict the diseases the user might be suffering from based on the symptoms given as input. Random forest model had the highest accuracy of 95.7% followed by K-Nearest Neighbor, Naïve Bayes Classifiers and Decision Tree having accuracies of 95.6%, 94.5% and 92.4% respectively.

### III. THE PROPOSED METHOD

#### A. Data Collection

It is a process where the required data is collected from various available resources and the necessary data is loaded to analyze and formulate patterns. This step involves acquiring the required datasets from the Kaggle repository and dividing the dataset into two datasets, i.e., training dataset and the testing dataset, as cited in [5]. The datasets consist of 132 features in total with 4920 records in the training dataset and 41 records in testing as in [5][14].

#### B. Data Preprocessing

It is a process that helps to transform raw data into insightful data that can be used for business decisions. The steps involved in Data Preprocessing are acquiring the dataset, importing the dataset and the required libraries, recognizing and handling the missing values as in [10][18], encoding the categorical data [12][18], splitting the dataset and feature scaling to obtain relevant data [18], which in return produces outcomes with better accuracy as cited in [5][12]. In addition to the existing datasets, as mentioned in [5][14], the datasets have been preprocessed by adding records in both training and testing datasets to improve the model's performance and for more accurate prediction as done in [10]. The proposed data set consists of 5000 training records and 54 testing records. Furthermore, an extra classification in the target feature has been added, indicating to the user that they are healthy when they do not suffer from any symptoms.

#### C. Working of the model

The initial step involves importing various libraries that support machine learning algorithms [13] such as NumPy, pandas, sklearn, statistics and tkinter into the proposed model. This step is followed by reading the training and testing datasets as in [5]. After the preprocessing stage, the next step involves splitting the training and testing datasets into feature and target attributes. Then, by using the sklearn module, import various machine learning classifiers to fit the training data into the models. Multiple machine learning classifiers like Decision Tree [1][5][8], Random Forest [1][2][6][7][14], K-Nearest Neighbor [4][5][9][11], Support Vector Machine [1][4][16], Naïve Bayes Classifier [1][4][10][11] and Logistic Regression [4][9] are used for training and prediction.

The predictions from the above models are stored in a global list (used to store multiple similar or dissimilar values), and the mode function present in the statistics library is applied to the list to obtain the most probable disease the person might be suffering.

The use of statistical mode as shown in Fig.1. in the proposed model helps in receiving the final prediction. The predictions received from individual machine learning algorithms are combined and placed in a global list. On applying the mode () function present in the statistics library, the most frequently predicted disease is returned.

$$\text{Final Prediction} = \text{Mode} (p_1, p_2, p_3, p_4, p_5, p_6)$$

$p_i$  - prediction of algorithm  $i$  (for  $i = 1, 2, \dots, 6$ )

- This result is thus considered the final prediction as shown in Fig.1. The final prediction resulting from the combined predictions of the classifiers [11] makes the model much more robust. It makes the



2581-4575

# International Journal For Recent Developments in Science & Technology



A Peer Reviewed Research Journal



prediction even more accurate and results better than the individual models cited in [5].

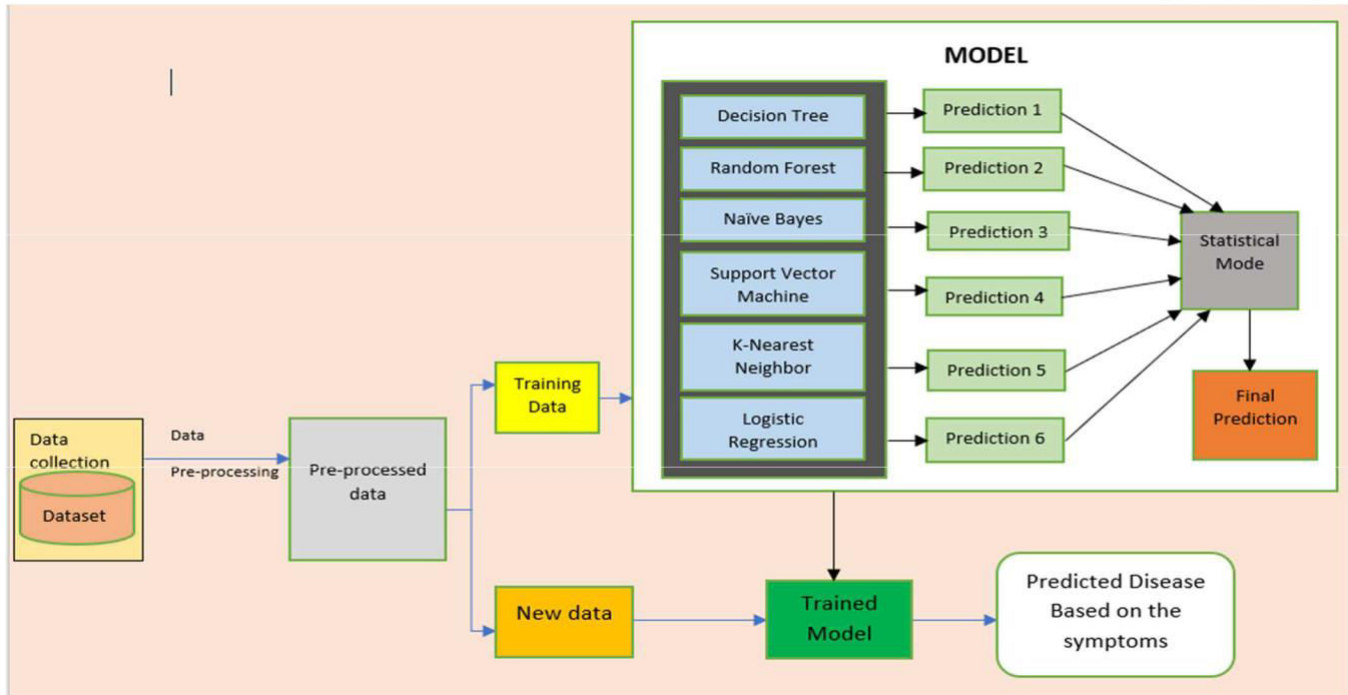


Fig. 1. System Architecture

#### D. New Data

When the user enters new data in the form of symptoms [11], the prediction is returned, indicating the most probable disease the person might be suffering from. This new data is stored in the database to improve the strength of the dataset, which aids in better prediction for future purposes.

#### E. Procedure for Implementation

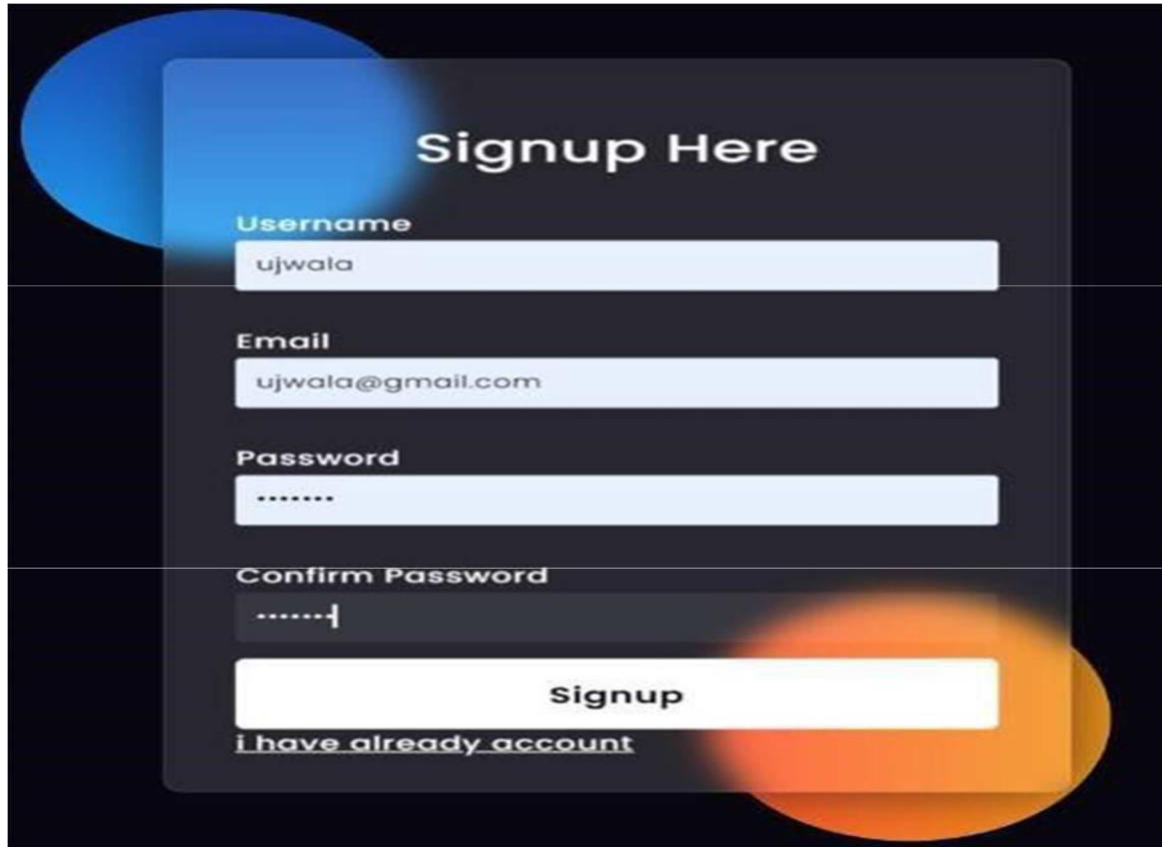
- Import the libraries such as NumPy, pandas, statistics, tkinter and sklearn into the proposed model.
- Import the datasets from Kaggle and perform various pre-processing techniques on the data.
- Split the training and testing datasets into feature and target attributes.
- Fit the training data into the models like Decision Tree [8], Logistic Regression [9], Random Forest [1][2][6][14], Naive Bayes Classifier, K-Nearest Neighbors and Support Vector Machine [8][16] for training and prediction.
- After training the models with relevant data, the statistical mode of the predictions made by the models is displayed, indicating the most probable disease that the person might be suffering from.
- In the webpage, enter the symptoms that the person is suffering from (that is new data is being entered).

- The most probable disease that the person might be suffering from based on the symptoms entered by the user is displayed by combining the predictions of the models.

## IV. RESULTS AND DISCUSSION

### A. Qualitative Results

- The project enhances the quickness of decision-making and provides a user-friendly GUI, for the user to enter the symptoms from the given list of symptoms and receive a prediction of the most probable disease that they might be suffering from based on the specific symptoms entered.
- The same functionality provided by the GUI is also provided to the users in the form of a webpage which was developed using the Django framework. It includes a signup page where the user can sign up, as shown in Fig. 2. and create a new account when they do not have an existing account.
- The user can click on 'I have already account' to be redirected to the login page, as shown in Fig. 3. On entering the login credentials, the user is redirected to the home page, as in Fig. 4., where he can enter the symptoms and get the prediction.



**Signup Here**

**Username**

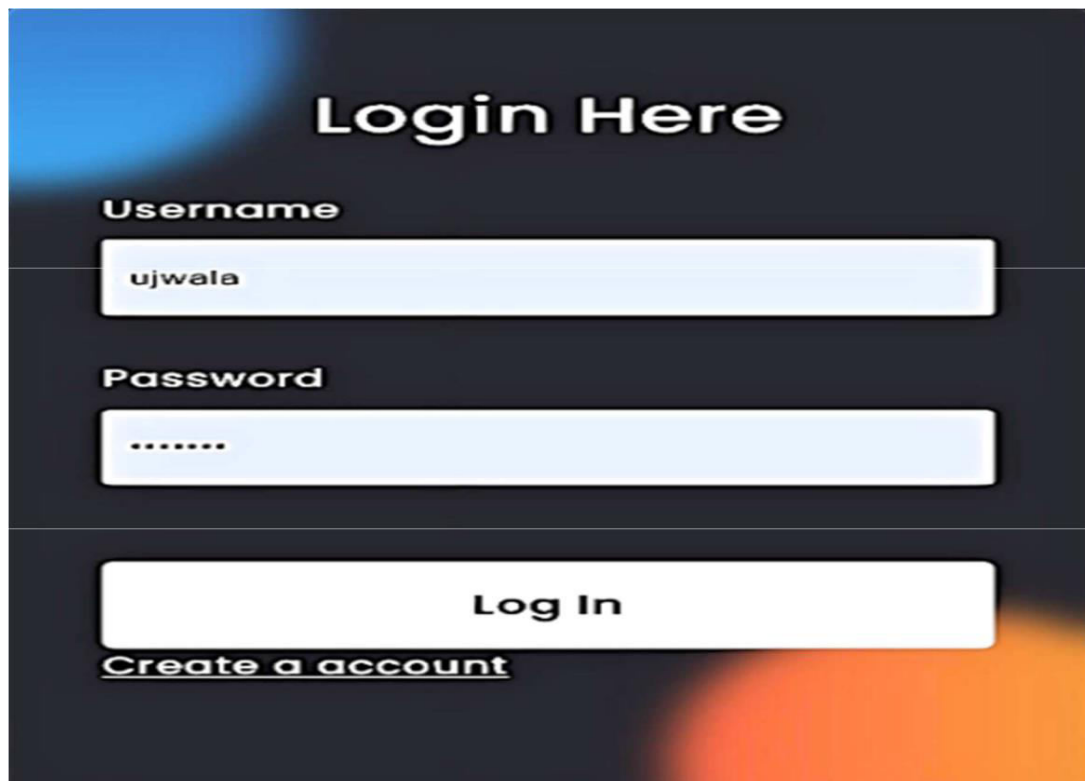
**Email**

**Password**

**Confirm Password**

[i have already account](#)

Fig. 2. Sign-up page



**Login Here**

**Username**

**Password**

[Create a account](#)

Fig. 3. Login page



2581-4575

# International Journal For Recent Developments in Science & Technology



A Peer Reviewed Research Journal



## B. Comparative Results

### Prediction

**Name of the Patient :**

**Symptom1 :**

**Symptom2 :**

**Symptom3 :**

**Symptom4 :**

**Symptom5 :**

---

**Predicted Disease :**

Fig. 4. Home Page

The proposed model output shown in Fig. 4. has been compared with the previous work as shown in Fig. 5. The previous work had a disease prediction system where each algorithm was considered as an individual model and the predictions made by the respective individual models were displayed. These models had individual accuracies and resulted in independent results.

When compared to previous work, the proposed model gives the user a simple and unambiguous user interface to get an accurate prediction on the most probable disease they might be suffering from. This prediction harnesses the combined power of six supervised machine learning algorithms. Thus, makes the prediction even more accurate and better than the results from individual models.

**Name of the Patient**  **Prediction 1**

**Symptom 1**  **Prediction 2**

**Symptom 2**  **Prediction 3**

**Symptom 3**  **Prediction 4**

**Symptom 4**  **Reset Inputs**

**Symptom 5**  **Exit System**

<b>DecisionTree</b>	<b>Common Cold</b>
<b>RandomForest</b>	<b>Allergy</b>
<b>NaiveBayes</b>	<b>Allergy</b>
<b>kNearestNeighbour</b>	<b>Allergy</b>

Fig. 5. Previous work output



2581-4575

# International Journal For Recent Developments in Science & Technology



A Peer Reviewed Research Journal





## C. Quantitative Results

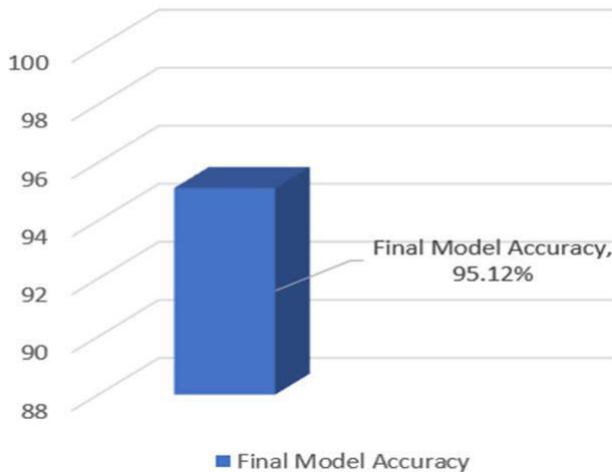


Fig. 6. Accuracy of proposed model

Accuracy of the final model developed by combining the six supervised machine learning algorithms is stated in Fig. 6.

## V. FUTURE SCOPE

- To add image data sets and use algorithms like CNN to predict diseases with more accuracy.
- To extend the model by using various classification techniques or algorithms apart from the already existing techniques to improve the accuracy. In future, more diseases can be added in the existing API.
- To improve the accuracy of prediction in order to decrease the mortality rate.

## VI. CONCLUSION

The study aims to predict the disease according to the symptoms selected by the patient by implementing a model which combines the predictions from various Supervised Machine Learning Classification Algorithms. In this study, six Machine Learning algorithms were used to obtain the final prediction and have achieved a mean accuracy of 95.1%. This shows exceptional refinement, higher accuracy and reliability than the preceding work. The model stores the data entered by the user along with the disease the patient is suffering from. These can be used as previous records and aid in future prognosis. This work involves a GUI for a smoother interaction with the model to improve the user experience and operability of the model. It aids hospitals, and healthcare organizations in making better decisions about how they provide care. The model performs a practical medical data analysis and predicts the disease before worsening the patient's health, hence allowing early patient care.

## REFERENCES

- [1] Mr. A. Rohith Naidu, Dr C K Gomathy: "The Prediction of Disease Using Machine Learning", International Journal of Scientific Research in Engineering and Management (IJSREM)
- [2] Prof. Suchita Wankhade, Rudra A. Godse, Karan A. Jagtap, Smita S. Gunjal, Mahamuni Neha S: "Multiple Disease Prediction Using Different Machine Learning Algorithms Comparatively", International
- [3] Keniya Rinkal, Ninad Mehendale, Aman, Khakharia, Mahesh Warang, Vruddhi Shah, Vrushabh Gada, Tirth Thaker, Manjalkar Ruchi: "Disease prediction from various symptoms using machine learning"
- [4] Megha Kamboj: "Heart Disease Prediction with Machine Learning Approaches", International Journal of Science and Research (IJSR) (2018)
- [5] Mr. Anap Pathak, Anuj Kumar: "A Machine Learning Model for Early Prediction of Multiple Diseases to Cure Lives", Turkish Journal of Computer and Mathematics Education, 2021
- [6] B. Prajna, Divya Mandem: "Multiple Disease Prediction System", International Journal of Innovative Research in Technology (IJIRT) (2021)
- [7] S.M.D. Jabeer, G. Chakravarthi: "Heart Disease Prediction Based on Machine Learning Techniques", International Journal of Innovative Research in Technology (IJIRT)
- [8] M.Lakshmi Narayana, O. Rama Praneeth Kumar, N.Sai Prasad, T. Naga Sampath, N Md Jubair Basha: "Chronic Disease Prediction Using gradient Boosting and KNN Classifier", International Journal of Innovative Research in Technology (IJIRT) October 2021
- [9] Mursal Furqan, Kanwal Awan, Hiba Rajput, Sanam Narejo, Adnan Ashraf: "Heart Disease Prediction Using Machine Learning Algorithms", International Conference on Computational Sciences and Technology December (2020)
- [10] Khurana Sarthak, Gupta Dr. Akhilesh Das, Jain Atishay, Bhasin Kunal, Kataria Shikhar, Aror Sunny: "Disease Prediction System", International Research Journal of Engineering and Technology, 6(5), 5178-5184.
- [11] Sarage Saurabh, Pingale Kedar, Kulkarni Vaibhav, Surwase Sushant, Karve Prof. Abhijeet: "Disease Prediction using Machine Learning", International Research Journal of Engineering and Technology, 6 (12), 2810-2813.
- [12] Rinal A, Chauhan Raj H, Naik Daksh N, Sagarkumar J, Halpati Patel, Prajapati Mr. A.D: "Disease Prediction using Machine Learning", International Research Journal of Engineering and Technology, 7(5), 2000-2002.
- [13] Gopi Battineni, Chinatalapudi Nalini, Francesco Amenta, Sagaro Getu Gamo: "Application Of Machine Learning Predictive Models in the Chronic Disease", International of PersonalisedMedicine, 10(21), 1-11.
- [14] Narendran. G, Pramoth Krishnan. T, Nivethitha. A: "Smart Disease Prediction Using Machine Learning", International Journal of Innovative Science and Research Technology, Volume 6, Issue 6, June – 2021
- [15] H. Patel, S. Patel, "Survey of data mining techniques used in the healthcare domain," Int. J. of Inform. Sci. and Tech., Vol. 6, pp. 53-60, March 2016.
- [16] K. Arumugam, Tatiana Gonzales-Yanac, Mohd Naved, Orlando Leiva-Chauca, Priyanka P. Shinde, Antonio Huaman-Osorio, Materials Today: Proceeding, August 2021
- [17] Computer Science, University of Benin: "Comparative Analysis of Machine Learning Algorithms for Heart Disease Prediction", International Journal of Scientific and Research Publications (IJSRP) 2021
- [18] Priya S. S., and Kumar, S. (2018). Chronic Kidney Disease Prediction Using Machine Learning. International Journal of Computer Science and Information Security (IJCSIS), 16(4)
- [19] T.V. Sriram, M.V. Rao, G.S. Narayana, D. Kaladhar, T.P.R. Vital, Intelligent parkinson disease prediction using machine learning algorithms, International Journal of Engineering and Innovative Technology (IJEIT) 3(3), 1568 (2013)
- [20] M. Maniruzzaman, M.J. Rahman, B. Ahammed, M.M. Abedin, Classification and prediction of diabetes disease using machine learning paradigm, Health Information Science and Systems 8(1), 7 (2020)