Multiple Disease Prediction System Using Machine Learning

AJIT KUMAR-20BCS7239 & SHUBHAM KUMAR-20BCS7174
Department of Computer Science and Engineering
Chandigarh University
Mohali, Punjab, India

AYUSH KUMAR PATHAK-20BCS4990
Department of Computer Science and Engineering
Chandigarh University
Mohali, Punjab, India

SWATI (E16448)
Department of Computer Science and Engineering
Chandigarh University
Mohali, Punjab, India

Abstract— In the face of increasingly serious health problems, "Multiple Infectious Diseases Using Machine Learning" aims to solve the fundamental problem: identifying and predicting various diseases to enable health management. Driven by the increasing amount of medical knowledge and the need for personalized and timely interventions, the aim is to transform disease prediction, prevention and management through the integration of higher education systems. The motivation behind this initiative stems from the increasing burden of various diseases and the need to move from rework to effective treatment strategies. Traditional medical models often fail to predict and prevent disease, increase healthcare costs, and impact patient outcomes. The project aims to analyze a wide range of medical data, including genetic information, lifestyle and medical history information, using the power of machine learning algorithms such as mixed methods and deep learning models.

This program is comprehensive and designed to predict the likelihood of many diseases based on a person's unique health status. Machine learning models, especially common methods such as random forest or gradient boosting, play an important role in prediction engines that use disparate data to provide accurate and personalized assessment.

The system integrates electronic medical records, genetic information and lifestyle data to provide recommendations for disease prevention and early intervention.

Through its intuitive and user-friendly interface, the initiative takes into account the principle that will enable doctors and individuals to make informed decisions about health risks and preventive measures in consumption. The system will enable timely intervention, reduce overall treatment burden and improve patient outcomes. By detecting dangerous diseases and promoting preventive measures, the program works to create a future where treatment is not only about rehabilitation but also prediction.

Introduction

In the changing healthcare environment, integration of technology has become imperative for accurate diagnosis and health management. Among these technologies, machine learning stands out as a powerful tool that can revolutionize the way we predict and prevent diseases. The

project, called "Multi-disease Prediction System Using Machine Learning", aims to predict and analyze the risk of three diseases using the capabilities of machine learning algorithms: diabetes, heart disease and Parkinson's disease.

The modern healthcare system faces problems in early diagnosis and timely intervention of various diseases that often cause health problems and increase the cost of treatment pains.

Traditional disease prediction methods are based on manual measurements and historical data, so their accuracy and efficiency are limited. The integration of machine learning provides flexibility by analyzing disparate data and eliminating complex patterns, ultimately improving predictive capabilities in healthcare.

Existing System

Current disease prediction methods lay the foundation for efficient and effective ways to solve problems related to predicting many diseases, including diabetes, heart disease, and Parkinson's disease. The system starts with the basic steps of data collection and aims to collect large and diverse medical records containing patient-related information and various medical information related to the disease.

Data collection is done carefully before processing to resolve missing results, negatives and ensure appropriate measurement parameters, thus making it essential for data quality and machine learning model learning. The basis of the method is in the training model, which uses various learning algorithms such as decision trees, random forests and artificial neural networks to learn patterns and relationships from previous literature. The system includes a robust phase selection model that uses performance metrics such as accuracy, precision, and recall to determine the best disease prediction. It is based on a rigorous evaluation model of independent test data and provides insight into the accuracy and reliability of the chosen model in predicting various diseases. The final step involves the development of a physician-friendly user interface to facilitate access to patient information and provide predictions for many diseases and ease of use for prediction. tool.

Background

The Overcoming the challenge of predicting many diseases (diabetes, heart disease, and Parkinson's disease) requires an integrated approach that includes data collection, modelling, early diagnosis, and community involvement. The following
guidelines describe the plan to use machine learning to create useful multi-pathology:

1. Data Collection: The first part of the system involves collecting large medical records containing patient information and various treatments for various diseases. This data will be used to train machine learning models.
2. Preliminary data: The collected data will be pre-processed to check for missing results, errors and perform statistical analysis. This process involves cleaning and preparing materials for modelling.
3. Training models: These tools involve training different learning machines such as decision trees, random forests, and neural networks on pre-existing data. The training model will be used for disease prediction.
4. Model selection: Measurements such as accuracy, precision, and recall will be used to compare the performance of different learning systems and select the best model for disease prediction.
5. Model evaluation: Selected models will be evaluated on separate data sets to evaluate their accuracy and reliability in predicting various diseases. This component of the system involves evaluating the model and evaluating its performance. 6. User Development: The final product of the system includes an improved user interface that allows physicians to access patient information and obtain predictions for many diseases. The interface will be designed to provide an easy-to-use disease prediction tool.

Literature Survey

The use of machine learning (ML) technology to predict various diseases has attracted much attention in recent years due to its potential to revolutionize healthcare by enabling early diagnosis and personalized treatment plans. Below is a research article on some of the important research in this field:

"Using electronic medical records to predict various diseases" (Nguyen et al., 2018) logistic regression, random forests, and slope boosting. to treat chronic diseases. The method is very accurate in predicting diseases such as diabetes, high blood pressure and heart diseases. People have developed a multi-target prediction model using machine learning that can predict the growth rate of various diseases. They use techniques such as multi-label classification and deep learning to predict disease based on patient data. Yang et al.

A method based on negative matrix factorization (NMF) is proposed for predicting multiple diseases using EHR data. They demonstrated the effectiveness of their method in predicting conditions such as diabetes, hypertension and hyperlipidemia. This work presents a multi-level deep learning task for simultaneous prediction of multiple diseases using EHR patient data. The model leverages common representation across organisms to improve prediction accuracy and generalization. A multitask study combining physical and self-reported data to predict various aspects of health, including the presence of various diseases, stress levels, and well-being is proposed. "Network" (Huang et al., 2020):

These studies demonstrate the various methods and techniques used to predict various diseases using machine learning techniques, from traditional methods to deep learning and integrating multiple methods. data to improve forecast accuracy and overall capabilities.

Techniques Used in Multiple Diseases Prediction using ML

Using machine learning to predict various diseases involves using different types of learning depending on the nature of the data, the complexity of the disease, and the expected outcomes. Here are some ideas:

Feature Engineering: This involves selecting raw materials and transforming them into features relevant to disease prediction. Features for medical records; demographic information, medical history, diagnoses, medical records, genetic information, lifestyle factors, etc. may contain. Classification is useful for tasks such as predicting whether a patient has a disease. High dimensional feature space. schema files, but generally require more data and computing resources. Techniques such as bagging (e.g. Random Forest), boosting (e.g. AdaBoost) or stacking can be used. It helps to understand disease subtypes or patient classification. Deep learning: Using multilayer neural networks to learn hierarchical representation from raw data. Convolutional neural networks (CNN) can be used to analyze medical images, while recurrent neural networks (RNN) are suitable for continuous data such as real-time medical data. Extract features from images or text files. Improving these small clinical data models can improve performance. Adapt and improve the performance of the translation model. AUC or precision recall curve.

Interpretation of models: In clinical settings, interpretable models such as decision trees or linear models may be preferred over black box models that are difficult to understand and accepted by physicians. Model validation and interpretation: Validation strategies such as cross-validation, bootstrapping, or external validation of independent data sets are important to evaluate the generalizability of the model. Interpretation techniques such as SHAP (Shapley Additive explanations) or LIME (Local Interpretable Model-agnostic Explanations) can help understand model predictions. It is anticipated that machine learning models can provide information for disease diagnosis, prognosis and treatment planning.

Proposed System

We propose methods using machine learning (ML) for disease prediction has several important characteristics and assumptions. Here's a high-level overview of how to do this:

Data Collection: Collect pertinent medical information, including patient demographics, medical history, symptoms, diagnosis, imaging studies, and other relevant information. This information can come from electronic health records (EHRs), quality control tools, patient surveys, and more.

Data preprocessing: Clean the collected data, resolve missing values, and perform the construction process to remove features relevant to disease prediction. This step may include data modelling, coding categorical variables, and other preliminary procedures. Custom selection helps reduce size and can improve model performance and interpretation. Dichotomous outcomes were classified and regression was used to predict continuous outcomes (such as violence). Common ML algorithms used for disease prediction include logistic regression, decision trees, random forests, support vector machines (SVM), and neural networks. This involves splitting the data into training and reference sets, tuning model hyperparameters, and evaluating model performance using appropriate metrics (e.g., accuracy, precision, regression, F1 score, area under the ROC curve). Model evaluation: Evaluate the performance of a training model using adversarial and/or disaggregated data. Models are evaluated on their ability to accurately predict the likelihood of disease without being overworked. This may include integrating the model with existing treatments or creating standalone

applications for doctors or patients. Updates with new or updated information to ensure it remains accurate and current. Protect patient privacy and confidentiality throughout data collection, prioritization and distribution processes. This connectivity should allow doctors or patients to access relevant information, view disease occurrences or outcomes, and interpret predictive models. Doctors and software engineers to ensure the efficiency, accuracy and usability of systems in healthcare.
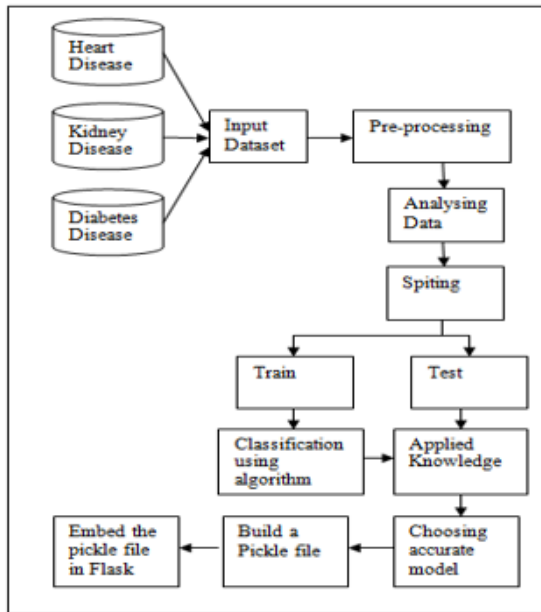

Fig. 1 Block Diagram

Figure: Block Diagram

Objectives

The aiming to predict many diseases using machine learning (ML), it is important to set clear goals to guide the development and evaluation of predictive models. Some of the main goals to consider are:

Accurate Disease Prediction: The main goal is to create an ML model that can predict the presence or absence of disease. According to patient information, there are many diseases. The model needs to achieve accuracy, precision, recall and F1 scores across all strains. The model must understand subtle patterns in the data that indicate the onset of various diseases, allowing for timely medical intervention and improving patient outcomes. Layer by layer. Machine learning models are expected to provide a risk score or probability estimate for each disease, allowing doctors to target high-risk patients for further evaluation or prevention. Factors determining treatment plans and their impact on prognosis are related to many diseases. Incorporate patient-specific factors such as demographics, genetics, lifestyle, and comorbidities into predictive models to optimize treatment decisions. ML models are expected to reveal interactions between many different treatments and provide insight into the coexistence of various diseases in human patients.

Model Interpretability: Ensure ML model predictions are interpretable and correctable. Ensure physicians understand the factors that contribute to prognosis, so they can understand the logic of the proposed model and make informed decisions. It generalizes well to unseen data and different patient populations. Analyze the performance model of external data to assess its reliability and validity in different clinical settings and geographic areas. Ethical and regulatory standards apply. Ensure patient privacy, confidentiality and integrity throughout the design, implementation and delivery phases. Evaluate its impact on clinical decision-making, patient outcomes, and utilization of healthcare resources to assess its true value and value in healthcare. Working together, researchers and clinicians can develop effective predictive models that will improve diagnosis, risk stratification, and personalized care.

Scope

The ability of machine learning (ML) to predict many diseases is immense and expands all aspects of medicine and disease management. Here is an overview of this study:

Disease Analysis: The main aim is to predict the presence or absence of various diseases based on the patient's data. This includes common diseases (such as diabetes, high blood pressure, heart disease) and rare diseases, including many conditions. It is a risk group for the development of many diseases. By identifying high-risk individuals, healthcare providers can implement preventive measures, early intervention, and screening to reduce disease and improve health. It is important to identify relationships and different patterns of variation. Machine learning models can reveal interactions between diseases, risk factors, and patient populations, providing insight into disease transmission and burden. Features for personal risk assessment and treatment planning. Predictive models can refine treatment strategies and improve patient outcomes by integrating patient-specific information such as genetics, lifestyle, medical history, and biomarkers. Assist doctors with diagnosis, prognosis assessment, and treatment options. By providing recommendations based on evidence and risk assessment, these

standards allow physicians to make informed decisions based on best practices and patient preferences. Research to assess disease diversity, measure health care disparities, and allocate resources efficiently. Predictive models can identify high-risk groups, prioritize public health interventions, and track the impact of preventive measures over time. Medical help is needed. This allows doctors to allocate the right staff, facilities and medical equipment to ensure that patients with various ailments receive timely care. Make new discoveries, develop hypotheses, and identify new biomarkers or disease entities. Using advanced analytical techniques, researchers can discover hidden patterns, biomarkers, and therapeutic targets that can help predict and manage diseases. The telemedicine and remote monitoring platform can realize remote disease monitoring, distribution, and decision support. This helps manage chronic conditions, reduces hospital admissions and allows patients to continue self-care. Provide information about organizations and levels of government. Using predictive analytics, policymakers can plan interventions, allocate funding, and implement policies to meet the complex health needs of multi-patient populations. The work to predict many diseases spans all clinical areas, from patient care to public health management, research and policy. By leveraging the power of machine learning and big data analytics, healthcare systems can increase the accuracy of disease prediction, improve patient outcomes, and rapidly improve the utilization of resources in medical care.

Methodology

Here's a way to predict many diseases using machine learning (ML):

Data collection and preprocessing:

Fill out a general information form with medical information and information about the disease you want to predict. This may include demographics, laboratory results, symptoms, and past diagnoses. This includes handling missing values, outliers, and inconsistencies. Techniques such as interpolation, scaling, and normalization may be necessary. Feature Engineering:

Analyze data to identify the most important features leading to disease prediction. The process selection process can help eliminate inconsistent features and improve performance standards. Model selection and training:

Selecting appropriate machine learning algorithms for disease prediction. Popular options include:

Support Vector Machines (SVM): Suitable for large data and can eliminate many errors. It can be applied to many types of scenarios. Prediction is useful for tasks that involve image analysis (such as diagnosing skin cancer). . Model evaluation and optimization:

Evaluate each model's performance using metrics such as accuracy, precision, recall, and F1 score. This helps measure how well the model predicts each disease. Different algorithms to tune the model or collect more data as needed. Deployment and usage:

Once you are satisfied with the performance of the model, deploy it to a valid application. This can be a web application or a mobile application where users can access information and get predictions about various diseases. significantly affects the accuracy of the model. Make sure your information is clear, accurate and representative of your target demographic. If interpretation is important, consider using models such as decision trees or rule-based methods. Recognize biases and take steps to reduce them. Please also emphasize that ML prediction is not a definitive diagnosis and should be used with clinical knowledge.

Algorithms Used

The Logistic Regression

Logistic regression analyzes the relationship between a categorical dependent variable and a group of independent (explanatory) variables. The name logistic regression is used when the variable has only two values, such as 0 and 1 or yes and no. Costs such as marriage, singleness, divorce or widowhood. Although the data type used for the variable in multiple regression is different, the application of the procedure is similar. Logistic regression competes with discriminant analysis as a method of examining response variables. Many statisticians consider logistic regression to be more than discriminant analysis and more suitable for modeling a variety of situations. This is because logistic regression does not assume that the independent variables are normally distributed as compared to discriminant analysis.

The program calculates binary logistic regression and multinomial logistic regression for numerical and categorical independent variables. Describes the regression equation including goodness of fit, variance, confidence limits, goodness of fit, and bias. Completes residual analysis, including

residual analysis and planning. It can be used as regression analysis to find the best regression model with the fewest independent variables. It provides a confidence interval for predicted values and provides a ROC curve to help determine the optimal cutoff for classification. It allows you to use the results by splitting unused rows during analysis.

Random Forest

Random Forest or Random Decision Forest is a common learning method for classification, regression and other tasks that works by generating

multiple decision trees at run time. For classification tasks, the results of random forest are classes selected from the majority of trees. For regression functions, return the mean or median estimate of a tree. Random decision forests correct the behaviour of decision trees that overfit their training. Random forests generally perform better than deterministic trees, but not less than gradient boosted trees. However, characteristics of the data may affect its performance.

The first decision random forest algorithm was developed by Tin Kam Ho [1] in 1995 using random subspace method. In Ho's formulation, this is an application of the classification of methods proposed by

" Random discriminant". Eugene Kleinberg.

SVM

In classification tasks, discriminative machine learning aims to find a discriminator based on independent and distributed (id) training dataset that can accurately predict the label of the newly acquired sample. Unlike generative machine learning, which must be calculated from relevant distributions, the exclusion distribution takes data point x and assigns it to one of several groups as part of the division of labour. Less powerful than the established method, which is generally used when anomalies need to be detected in the forecast, the discrimination method requires less expense and information, especially when it is required for many specific locations and only later. From a geometric perspective, working into classes is equivalent to finding the equal area that is better than the difference between classes at a given location.

SVM is a discriminant algorithm that always returns similar to the negative hyperplane as it solves the optimization problem, unlike Genetic Algorithms (GA) or Perceptron where both algorithms are widely used for classification in machine learning. For sensors, resolution depends on threshold and termination. Training for the custom kernel, which transforms the material from

the input domain to the custom domain, returns the SVM random model specific to the training process, while the perceptron and GA classifier model differ starting from each training session. The goal of genetic algorithms and detectors is simply to minimize errors during training; which will evolve into many hyperplanes to meet the requirements. norms.

```
input_data = (197.07600,206.89600,192.05500,0.00289,0.00001,0.00166,0.00168,0.00498,0.01098,0.09700,0.00563,0.00680,0.00802,0.016
input_data_as_numpy_array = np.asarray(input_data)
input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)
std_data = scaler.transform(input_data_reshaped)

prediction = model.predict(std_data)
print(prediction)


if (prediction[0] == 0):
  print("The Person does not have Parkinsons Disease")

else:
  print("The Person has Parkinsons")


[0]
The Person does not have Parkinsons Disease

C:\Users\ajitm\Downloads\Anaconda\Lib\site-packages\sklearn\base.py:464: UserWarning: X does not have valid feature names, but
StandardScaler was fitted with feature names
  warnings.warn(
```

Figure: Real Time

APPROACH

The Using machine learning to predict multiple diseases requires a more complex approach than predicting a single disease. Here is a method to overcome this challenge:

Data collection: Collect medical data covering the various diseases you are interested in. This information should include the patient's demographic information, medical history, symptoms, laboratory findings, clinical findings, genetic information (if available), and other treatment information. To build a good predictive model, make sure the dataset covers a wide range of diseases. This may include data structure, encoding of categorical variables, handling of unequal data (if necessary), and well-designed functions to extract meaning from raw data. Meanwhile, use special selection or extraction methods to determine the most data to predict various diseases. Techniques such as principal component analysis (PCA), importance ranking, or feature selection can be used to reduce dimensionality and improve model performance. - Work sharing. Common algorithms for multivariate classification include random forest, support vector machine (SVM), gradient boosting algorithms (such as XG-Boost, Light-GBM), and deep learning models such as connected neural network (CNN) or neural networks (RNN).

Training model: Trains the selected ML model using previous data. Use techniques such as using

crosshairs to optimize hyperparameters and make the model more efficient. Given the complexity of predicting multiple diseases, collaborative studies may be useful in combining predictions from multiple baseline models. value, F1 score and confusion matrix analysis. Note the model's performance across different diseases to ensure that predictions are equal. Techniques such as factor analysis, SHAP (Shapley Additive explanations) or LIME (local interpretation of mean uncertainty) can provide distressing insights into the decision model. Production environment to ensure suitability for processing or use. Create user-friendly interfaces for clinicians to efficiently access patient data and interpret predictive models. To maintain its effectiveness, accuracy and timeliness. Provide feedback from doctors to adjust the model and resolve any issues that arise. Justice system and administration of justice informed clinical decisions.

Result

The results of using machine learning (ML) for many disease predictions can vary depending on factors such as data quality, algorithm selection, engineering methods, and testing models. The results are typically as follows:

Model Performance Metrics: The performance of an ML model can be measured using a variety of metrics such as accuracy, precision, recall, F1 score, and area under the receiver operating curve (AUC). -ROC). These measurements provide insight into how well the model predicts the presence or absence of various diseases in various groups. It helps to see the real advantages, disadvantages, drawbacks and drawbacks, allowing a deeper understanding of the advantages and disadvantages of the model. It is especially important in class imbalance. Examining metrics such as precision, recall, and F1 score for each disease group can indicate whether the model performs well in each group or whether there are certain diseases for which accurate predictions are difficult to make. Factors that are more likely to predict may provide a better understanding of the relationship between clinical variables and disease outcomes. Factor analysis helps identify the most important factors that drive the forecast model. Doctors know how well the model predicts outcomes. Interpretive standards are important to build trust and facilitate clinical decision-making. additional cost. The fact that the ML model performs better than the traditional method indicates that it can improve the accuracy of

diagnosis and patient care. Evaluating its effectiveness in assisting physicians with diagnosis, risk stratification, treatment planning, and patient management is important to confirm its effectiveness and adoption in treatment. By using machine learning to predict outcomes of various diseases, including various performance measures and assessments, researchers and physicians can understand the model's potential and limitations, ultimately improving patient and treatment outcomes.

Conclusion

Although extensive research has been conducted on the classification of skin diseases in humans, little research has been done on skin diseases affecting animals, especially dermatoses in cattle. This study addresses this gap by presenting a method that uses image processing and machine learning techniques to distinguish animals with infected skin from infected skin that is always there. The developed technology demonstrated a high degree of accuracy in identifying skin diseases, providing an effective method for investigating animal diseases.
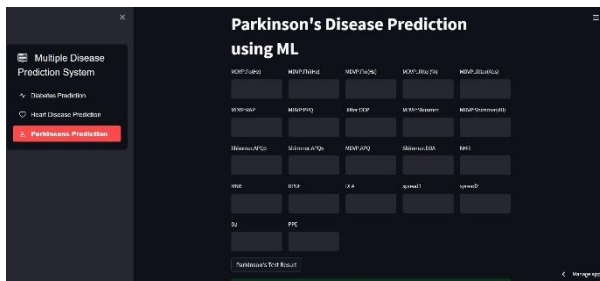
Figure: Outputs

References

Li 1. H. EL Massari, S. Mhammedi, Z. Sabouri, and N. Gherabi, "Ontology-Based Machine Learning to Predict Diabetes Patients," in Advances in Information, Communication and Cybersecurity, Cham, 2022, pp. 437–445. doi: 10.1007/978- 3-030-91738-8_40.

2. F. Alaa Khaleel and A. M. Al-Bakry, "Diagnosis of diabetes using machine learning algorithms," Mater. Today Proc., Jul. 2021, doi: 10.1016/j.matpr.2021.07.196.

3. . J. J. Khanam and S. Y. Foo, "A comparison of machine learning algorithms for diabetes prediction," ICT Express, vol. 7, no. 4, pp. 432–439, Dec. 2021, doi: 10.1016/j.icte.2021.02.004.

4. P. Cıhan and H. Coşkun, "Performance Comparison of Machine Learning Models for Diabetes Prediction," in 2021 29th Signal Processing and Communications Applications Conference (SIU), Jun. 2021, pp. 1–4. doi: 10.1109/SIU53274.2021.9477824.

5. S. Barik, S. Mohanty, S. Mohanty, and D. Singh, "Analysis of Prediction Accuracy of Diabetes Using Classifier and Hybrid Machine Learning Techniques," in Intelligent and Cloud Computing, Singapore, 2021, pp. 399–409. doi: 10.1007/978-981-15-6202-0_41.

6. Aditi Gavhane, Gouthami Kokkula, Isha Panday, Prof. Kailash Devadkar, "Prediction of Heart Disease using Machine Learning", Proceedings of the 2nd International conference on Electronics, Communication and Aerospace Technology (ICECA), 2021.

7. Md. Redone Hassan et al, "A Knowledge Base Data Mining based on Parkinson's Disease" International Conference on System Modelling & Advancement in Research Trends, 2019.

8. Santhana Krishnan J and Geetha S, "Prediction of Heart Disease using Machine Learning Algorithms" ICIICT, 2019.

9. Senthil kumar mohan, chandrasegar thirumalai and Gautam Srivastva, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques" IEEE Access 2019.

10. M. Nikhil Kumar, K. V. S. Koushik, K. Deepak, "Prediction of Heart Diseases Using Data Mining and Machine Learning Algorithms and Tools" International Journal of Scientific Research in Computer Science, Engineering and Information Technology, IJSRCSEIT 2019.

11. Dr. Anupam Bhatia and Raunak Sulekh, "Predictive Model for Parkinson's Disease through Naive Bayes Classification" International Journal of Computer Science & Communication vol. 9, Dec. 2017, pp. 194- 202, Sept 2017 - March 2018.