https://doi.org/10.5281/zenodo.14066308

# INTELLIGENT DISEASE PREDICTION SYSTEMS WITH MACHINE LEARNING ALGORITHMS

*[1]M.Mounika, [2]Ragula Vamshikrishna*
*[1]Assistant Professor,[2]MCA Student*
*Depatment Of MCA*
*Sree Chaitanya College of Engineering, Karimnagar*

## ABSTRACT

Predictive modelling is the foundation of the disease prediction system, which uses the user's symptoms as input to determine the user's illness. The user's symptoms are analysed by the algorithm, which then outputs the likelihood of the illness. Implementing the Decision Tree, Random Forest, and Naïve Bayer algorithms is how disease prediction is done. These algorithms determine the likelihood of the illness. Accurate analysis of medical data improves patient care and early illness diagnosis as big data grows in the biomedical and healthcare sectors. This method makes use of machine learning algorithms to forecast illnesses. Currently, the decision tree method will be used for illness prediction. We are employing the Random Forest and Naïve Bayer algorithms to improve.

## I. INTRODUCTION

Machine learning is programming computers to optimize a performance using example data or past data. Machine learning is the study of computer systems that learn from data and experience. Machine learning algorithm has two passes: Training, Testing. Prediction of a disease by using patient’s symptoms and machine learning technology. Machine Learning technology gives a good platform in medical field, so that healthcare issues can be solved efficiently. We are applying machine learning to maintained complete hospitals' data Machine learning technology which allows building models to get quickly analyze data and deliver results faster, with the use of machine learning technology doctors can make good decision for patient diagnoses and treatment options, which leads to improvement of patient healthcare services. Healthcare is the most prime example of how machine learning is use in medical field. To improve the accuracy from a large data, the existing work will be done on unstructured and textual data. For prediction of diseases, the existing will be done on decision Tree algorithm. For improving, we are using Random Forest Algorithm and Naïve Bayer Algorithm.

## DISEASE PREDICTION

Disease Prediction using Machine Learning is a system which predicts the disease based on the information provided by the user. It also predicts the disease of the patient or the user based on the information or the symptoms he/she enter into the system and provides the accurate results based on that information. If the patient is not much serious and the user just wants to know the type of disease, he/she has been through. It is a system which provides the user the tips and tricks to maintain the health system of the user and it provides a way to find out the disease using this prediction. Now a day’s

https://doi.org/10.5281/zenodo.14066308

health industry plays major role in curing the diseases of the patients so this is also some kind of help for the health industry to tell the user and also it is useful for the user in case he/she doesn’t want to go to the hospital or any other clinics, so just by entering the symptoms and all other useful information the user can get to know the disease he/she is suffering from and the health industry can also get benefit from this system by just asking the symptoms from the user and entering in the system and in just few seconds they can tell the exact and up to some extent the accurate diseases. This DPUML is previously done by many other organizations but our intention is to make it different and beneficial for the users who are using this system. This Disease Prediction Using Machine Learning is completely done with the help of Machine Learning and Python Programming language with Tkinter Interface for it and also using the dataset that is available previously by the hospitals using that we will predict the disease. Now a day’s doctors are adopting many scientific technologies and methodology for both identification and diagnosing not only common disease, but also many fatal diseases. The successful treatment is always attributed by right and accurate diagnosis. Doctors may sometimes fail to take accurate decisions while diagnosing the disease of a patient, therefore disease prediction systems which use machine learning algorithms assist in such cases to get accurate results. The project disease prediction using machine learning is developed to overcome general disease in earlier stages as we all know in competitive environment of economic development the mankind has involved so much that he/she is not concerned about health according to research there are 40% peoples how ignores about general disease which leads to harmful disease later. The main reason of ignorance is laziness to consult a doctor and time concern the peoples have involved themselves so much that they have no time to take an appointment and consult the doctor which later results into fatal disease. According to research there are 70% peoples in India suffers from general disease and 25% of peoples face death due to early ignorance the main motive to develop this project is that a user can sit at their convenient place and have a check-up of their health the UI is designed in such a simple way that everyone can easily operate on it and can have a check-up.

## PROBLEM DEFINITION

Now a day’s in Health Industry there are various problems related to machines or devices which will give wrong or unaccepted results, so to avoid those results and get the correct and desired results we are building a program or project which will give the accurate predictions based on information provided by the user and also based on the datasets that are available in that machine. The health industry in information yet and knowledge poor and this industry is very vast industry which has lot of work to be done. So, with the help of all those algorithms, techniques and methodologies we have done this project which will help the peoples who are in the need. So the problem here is that many people goes to

https://doi.org/10.5281/zenodo.14066308

hospitals or clinic to know how is their health and how much they are improving in the given days, but they have to travel to get to know there answers and sometimes the patients may or may not get the results based on various factors such as doctor might be on leave or some whether problem so he might not have come to the hospital and many more reasons will be there so to avoid all those reasons and confusion we are making a project which will help all those person's and all the patients who are in need to know the condition of their health, and at sometimes if the person has been observing few symptoms and he/she is not sure about the disease he/she is encountered with so this will lead to various diseases in future. So, to avoid that and get to know the disease in early stages of the symptoms this disease prediction will help a lot to the various people's ranging from children to teenagers to adults and also the senior citizens.

## PROJECT PURPOSE

The purpose of making this project called "Disease Prediction Using Machine Learning" is to predict the accurate disease of the patient using all their general information's and also the symptoms. Using this information, there we will compare with our previous datasets of the patients and predicts the disease of the patient he/she is been through. If this Prediction is done at the early stages of the disease with the help of this project and all other necessary measure the disease can be cured and in general this prediction system can also be very useful in health industry. If health industry adopts this project then the work of the doctors can be reduced and they can easily predict the disease of the patient. The general purpose of this Disease prediction is to provide prediction for the various and generally occurring diseases that when unchecked and sometimes ignored can turns into fatal disease and cause lot of problem to the patient and as well as their family members. This system will predict the most possible disease based on the symptoms. The health industry in information yet and knowledge poor and this industry is very vast industry which has lot of work to be done. So, with the help of all those algorithms, techniques and methodologies we have done this project which will help the peoples who are in the need.

## OBJECTIVE

There is a need to study and make a system which will make it easy for an end user to predict the chronic diseases without visiting physician or doctor for diagnosis. To detect the Various Diseases through the examining Symptoms of patient's using different techniques of Machine Learning Models. The Predictions Accuracy will increase using advance Machine Learning.

## II. LITERATURE SURVEY

[1] In 2010, Apache Hadoop sharp big data as "datasets which could not be apprehended, succeeded, and managed by general computers within an okay scope." On the basis of this definition, in May 2011, McKinsey & Company, a global accessing help said Big Data as the next edge for improvement, war, and yield. Big data shall callous such datasets which could not be attained, succeeded and

https://doi.org/10.5281/zenodo.14066308

stored by standard database software. This classification includes two associations: First, datasets dimensions that obey to the usual of big data are shifting, and may cultivate over time or with scientific developments. Second, datasets measurements that adapt to the ordinary of big data in unalike submissions contrast from each other.

[2] Clinical data recounting the phenotypes and dealing of patients denotes an underused data font that has much bigger research likely than is currently grasped. Mining of electrical health records has the facility to form a new patient-stratification doctrines and for tight fitting unknown disease links. Mixing EHR data with genetic data will also give a more kind of genotype-phenotype affairs. However, a wide series of permitted, ethical, and methodological reasons presently hold back the organized confession of these data in electrical health histories and their excavating. Here, it consider the likely for furthering medical examination and experimental care using EHR data and the tasks that must be dazed before this is a truth.

[3] The medical resources of many countries are limited. For example, in China, the growth of medical resources is not balanced that 80% people are living in areas with inadequate medical resources while 80% medical resources are allocated at the big cities. Construction of big health application system by successfully mixing medical health resources using smart depots, health Internet of Things (IoT), big data and cloud computing is the vital way to resolve the above difficulties. Big health is a talented industry, which is characterized by people-center, managing a person's health from birth to decease, from anticipation to rehabilitation and involving industry from administration to market. The field of big health covers health goods field (including the drugs, medical devices, elder goods), health service field (including medical services, income services, mobile healthcare), fitness real estate field (including pension, healthcare) and health finance field (including health protection and other financial products).

[4] Chinese herbal products (CHPs) are commonly developed for patients with hyperlipidemia in traditional Chinese medicine (TCM). Since hyperlipidemia and connected sickness are public topics worldwide, this training discovered the drug shapes and occurrences of CHPs for giving patients with hyperlipidemia. Traditional Chinese medicine (TCM) has become common as a healing for central indicators in patients with hyperlipidemia. This drill likely to study the treatment patterns of TCM for patients with hyperlipidemia. The study population was recruited from a random-sampled troop of 1,000,000 folks from the National Wellbeing Insurance Exploration Record between. It recognized 30,784 fatality visits linked with hyperlipidemia judgment and collected these medical records. Overtone rules of facts withdrawal were led to moveable the co-prescription plans for Chinese herbal products (CHPs).

https://doi.org/10.5281/zenodo.14066308

[5] In this paper, it witness the use of recurrent neural networks (RNNs) with the situation of search-based operational publicity. It practice RNNs to map equally queries and ads to real valued vectors, by means of which the significance of a given (query, ad) couple can be simply calculated. On upper of the recurrent neural networks, it familiarize a novel consideration network, which studies to assign attention scores to different word locations according to their intent importance (hence the name Deep Intent). Later by this method, the path output of a arrangement is computed by a weighted sum of the hidden states of the RNN at each word according their attention scores. The system achieve end-to-end exercise of together the RNN and attention system below the guidance of user click logs. These worker click logs are sampled from a commercial search engine. It demonstrate that in most cases the attention network improves the quality of learned vector representations, evaluated by AUC on a physically labeled dataset. And furthermore, it highlight the effectiveness of the learned attention nicks from two aspects as: query rewriting and a modified BM25 metric. The system illustrate that using the learned attention scores, one will be able to produce sub-queries that would be of better qualities than those of the state-of-the-art methods. In count, by regulating the term occurrence with the care scores in a normal BM25 formula, one is bright to improve its performance evaluated by AUC.

[6] Abstract Traditional wearable devices have various drawbacks, such as uncomfortableness for long-term wearing, and insufficient accuracy, etc. Thus, health monitoring through traditional wearable devices is hard to be sustainable. In order to obtain and manage healthcare big data by sustainable health nursing, the system design “Smart Clothing”, enabling unobtrusive collection of various physiological indicators of human body. To offer persistent cleverness for smart clothing erection, mobile healthcare cloud stand is constructed by the usage of mobile internet, cloud computing and big data analytics. This paper announces design facts, key tools and applied implementation methods of smart dress system. Typical claims powered by smart clothing and big data clouds are presented, such as medical backup response, emotion care, disease diagnosis, and real-time tangible interaction.

[7] In this it extant a new deep learning manner Bi-CNN-MI for paraphrase identification (PI). Created on the vision that PI needs associating two sentences on many heights of granularity, it learn multigranular decree images using convolutional neural network (CNN) and model boundary features at each level. These topographies are then the input to a logistic classifier for PI. All limits of the model (for embeddings, convolution and classification) are straight optimized for PI. To address the lack of training data, the system pretrain the network in a novel method using a language modeling task. Results on the MSRP corpus surpass that of earlier NN competitors.

https://doi.org/10.5281/zenodo.14066308

[8]Does the estimate of lung cancer using the double dispensation system. The image dispensation system is familiarized into the double for early prophecy. The challenging in this progression is recognition of tiny nodes which comprehends early cancer finding. The unstipulated knobs in lungs can be spotted using ridge recognition algorithm.

[9] It proposed a system that integrates different datum such as gene information, DNA methylation, and miRNA. In this paper, the model has combined multiple kernel learning methods and dimensionality reduction.

[10] On the available data mining algorithms to classify the data and extract the knowledge from it. It discusses about the difficulties in classification, segmentation, extraction and selection. It compares the different algorithms like Support Vector Machine, Naïve Bayesian classification, Rough set theory, Decision Tree.

## III. SYSTEM ANALYSIS AND DESIGN

### EXISTING SYSTEM

The system predicts the chronic diseases which are for particular region and for the particular community. The Prediction of Diseases is done only for particular diseases. In this System, Big Data & decision Tree algorithm is used for Diseases risk prediction. Existing paper, we streamline machine learning algorithms for effective prediction of chronic disease outbreak in disease-frequent communities.

### PROPOSED SYSTEM

This system is used to predict most of the chronic diseases. It accepts the structured and textual type of data as input to the machine learning model. This system is used by end users. System will predict disease on the basis of symptoms. This system uses Machine Learning Technology. For predicting diseases Random Forest Algorithm and Naïve Bayer Algorithm are used.

## IV. SYSTEM DESIGN

### SYSTEM ARCHITECTURE

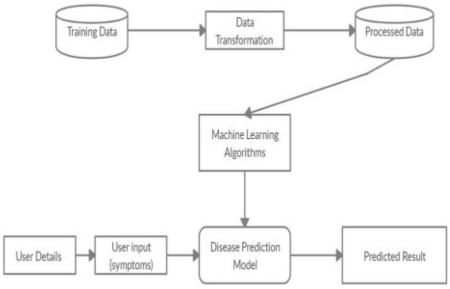


Fig: System Architecture

## V. SYSTEM IMPLEMENTATION

### MODULES

1. Data per processing
2. Apply Machine Learning Program
3. Predict Disease Based

### MODULE DESCRIPTION

### DATA PER PROCESSING

- Data collection and dataset preparation This will involve collection of medical information from various sources like hospitals, then pre-processing is applied on dataset which will remove all the unnecessary data and extract important features from data.
- Training and experimentation on datasets The Disease Prediction model will be trained on the dataset of diseases to do the prediction accurately and produce Confusion matrix.

### Apply Machine Learning Program

- In this project 3 different algorithms were used:

https://doi.org/10.5281/zenodo.14066308

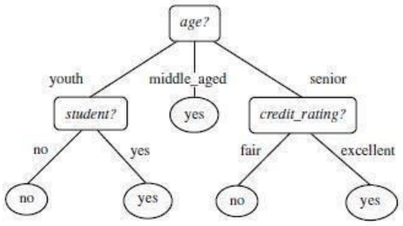
Decision Tree Algorithm

A. Random Forest Algorithm
B. Naïve Bayes Algorithm

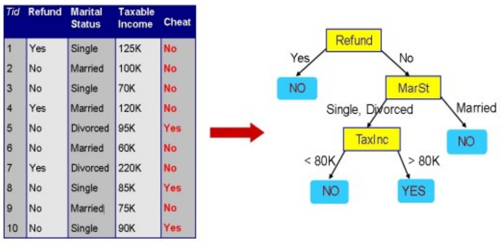
- Deployment and analysis on real life scenario the trained and tested prediction model will be deployed in a real-life scenario made by the human experts & will be leveraged for further improvement in the methodology.
- The working and basic explanation of those 3 algorithms Random Forest, Decision Tree and Naïve Bayes is given below.

**Decision tree algorithm**

Decision tree induction is the learning of decision trees from class-labelled training tuples. A decision tree is a flowchart-like tree structure.



- Decision tree induction is a non-parametric approach for building classification models.
- Finding an optimal decision tree is an NP-complete problem
- Techniques developed for constructing decision trees are computationally inexpensive, making it possible to construct models even when the training set size is very large.
- Decision trees, especially smaller-sized trees, are relatively easy to interpret.
- Decision tree provide an expressive representation for learning discrete-valued functions.
- Decision tree algorithms are quite robust to the presence of noise, especially when methods for avoiding overfitting.



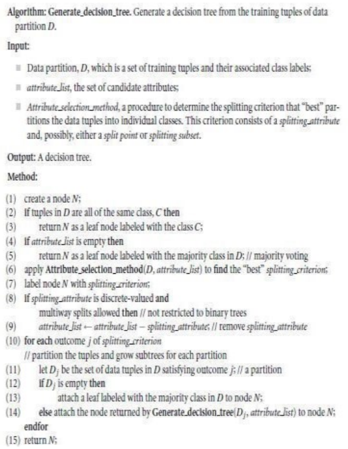
| Tid | Refund | Marital Status | Taxable Income | Cheat |
|---|---|---|---|---|
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 95K | Yes |
| 6 | No | Married | 60K | No |
| 7 | Yes | Divorced | 220K | No |
| 8 | No | Single | 85K | Yes |
| 9 | No | Married | 75K | No |
| 10 | No | Single | 90K | Yes |

Training Data

Model: Decision Tree

- The presence of redundant attributes does not adversely affect the accuracy of decision tree.
- The construction of decision tree classifiers does not require any domain knowledge or parameter setting, and therefore I appropriate for exploratory knowledge discovery. Decision trees can handle high dimensional data.
- Their representation of acquired knowledge in tree form is intuitive and generally easy to assimilate by humans.
- The learning and classification steps of decision tree induction are simple and fast.
- In general, decision tree classifiers have good accuracy.
- Decision tree induction algorithm shave been used for classification in many application areas, such as medicine, manufacturing and production, financial analysis, astronomy, and molecular biology.

https://doi.org/10.5281/zenodo.14066308



**Algorithm: Generate_decision_tree.** Generate a decision tree from the training tuples of data partition *D*.

**Input:**

- Data partition, *D*, which is a set of training tuples and their associated class labels;
- *attribute_list*, the set of candidate attributes;
- *Attribute_selection_method*, a procedure to determine the splitting criterion that "best" partitions the data tuples into individual classes. This criterion consists of a *splitting_attribute* and, possibly, either a *split point* or *splitting subset*.

**Output:** A decision tree.

**Method:**

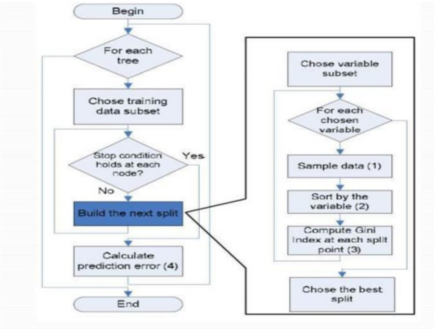
```
create a node N;
if tuples in D are all of the same class, C then
    return N as a leaf node labeled with the class C;
if attribute_list is empty then
    return N as a leaf node labeled with the majority class in D; // majority voting
apply Attribute_selection_method(D, attribute_list) to find the "best" splitting_criterion;
label node N with splitting_criterion;
if splitting_attribute is discrete-valued and
    multiway splits allowed then // not restricted to binary trees
    attribute_list ← attribute_list − splitting_attribute; // remove splitting_attribute
for each outcome j of splitting_criterion
// partition the tuples and grow subtrees for each partition
    let D_j be the set of data tuples in D satisfying outcome j; // a partition
    if D_j is empty then
        attach a leaf labeled with the majority class in D to node N;
    else attach the node returned by Generate_decision_tree(D_j, attribute_list) to node N;
endfor
return N;
```

**Fig: Decision Tree Algorithm**

## Random Forest Algorithm

It is an ensemble classifier using many decision trees models; it can be used for regression as well as classification.

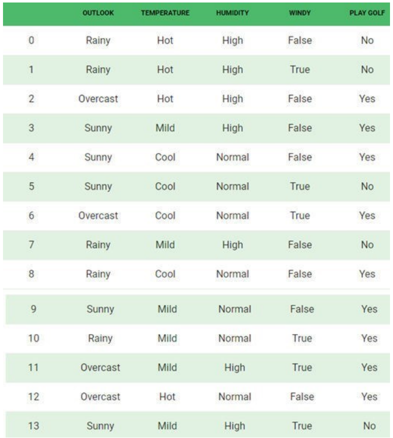
- Accuracy and variable importance information can be provided with the results.
- A random forest is the classifier consisting of a collection of tree structured classifiers k, where the k is independently, identically distributed random trees and each random tree consist of the unit of vote for classification of input.
- Random forest uses the Gini index for the classification and determining the final class in each tree.
- The final class of each tree is aggregated and voted by the weighted values to construct the final classifier.
- The working of random forest is, A random seed is chosen which pulls out at a random, a collection of samples from the training datasets while maintaining the class distribution.



## Naïve Bayes Algorithm

- It is used to predict the categorical class labels.
- It classifies the class data based on the training set and the values in a classifying attribute and uses it in classifying new data.
- It is a two-step process Model Construction and Model Usage.
- This Bayes theorem is named after Thomas Bayes and it is statistical method for classification and supervised learning method.
- It can solve both categorical and continuous values attributes.
- Bayes theorem finds the probability of an event occurring given the probability of another event that has already occurred. Bayes theorem is stated mathematically as the following equation.
- $P(A/B) = P(B|A)\ P(A)/P(B)$
- Below is the example how this algorithm/theorem works with the dataset.

https://doi.org/10.5281/zenodo.14066308

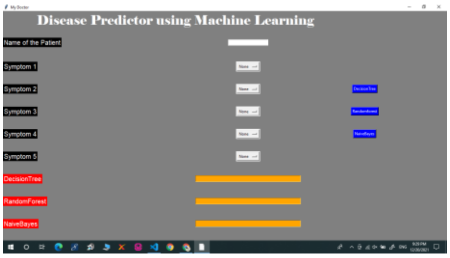


| | OUTLOOK | TEMPERATURE | HUMIDITY | WINDY | PLAY GOLF |
|---|---|---|---|---|---|
| 0 | Rainy | Hot | High | False | No |
| 1 | Rainy | Hot | High | True | No |
| 2 | Overcast | Hot | High | False | Yes |
| 3 | Sunny | Mild | High | False | Yes |
| 4 | Sunny | Cool | Normal | False | Yes |
| 5 | Sunny | Cool | Normal | True | No |
| 6 | Overcast | Cool | Normal | True | Yes |
| 7 | Rainy | Mild | High | False | No |
| 8 | Rainy | Cool | Normal | False | Yes |
| 9 | Sunny | Mild | Normal | False | Yes |
| 10 | Rainy | Mild | Normal | True | Yes |
| 11 | Overcast | Mild | High | True | Yes |
| 12 | Overcast | Hot | Normal | False | Yes |
| 13 | Sunny | Mild | High | True | No |

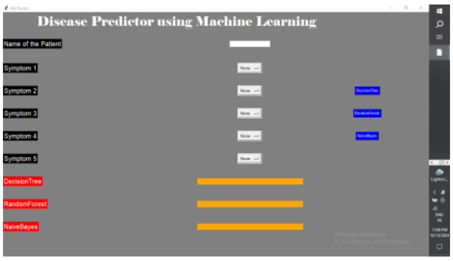
## PREDICT DISEASE BASED

- The given dataset is divided into two parts namely feature matrix and response vector.
- Feature matrix contains all the vectors means rows of the dataset in which each vector consists of the values of dependent features. In the above dataset features are outlook, temperature, humidity and windy.
- Response vector consist of values of class variables for each row of feature matrix. In the above dataset the class variable name is play golf.
- The fundamental naïve based assumption is that each feature makes an independent and equal contribution to the outcome.

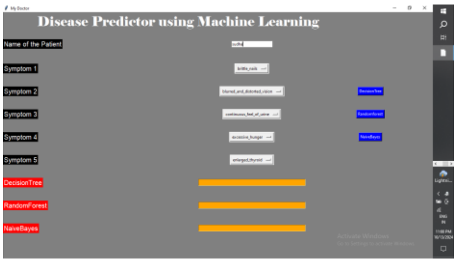
## VI. SCREEN SHOTS



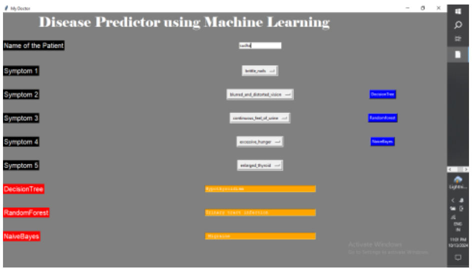
Symptom 1: None
Symptom 2: None
Symptom 3: None
Symptom 4: None
Symptom 5: None



Symptom 1: None
Symptom 2: None
Symptom 3: None
Symptom 4: None
Symptom 5: None



Name of the Patient: sudha
Symptom 1: brittle_nails
Symptom 2: blurred_and_distorted_vision
Symptom 3: continuous_feel_of_urine
Symptom 4: excessive_hunger
Symptom 5: enlarged_thyroid



Name of the Patient: sudha
Symptom 1: brittle_nails
Symptom 2: blurred_and_distorted_vision
Symptom 3: continuous_feel_of_urine
Symptom 4: excessive_hunger
Symptom 5: enlarged_thyroid
DecisionTree: Hypothyroidism
RandomForest: Urinary tract infection
NaiveBayes: Migraine

## VII. CONCLUSION

I will sum up by stating that this project, which uses machine learning to predict diseases, is very helpful in everyone's daily life. However, it is especially significant for the healthcare industry, as they are the ones who use these systems on a daily basis to predict the diseases of patients based on their general information and symptoms. If the health sector takes up this concept, doctors' workloads would be lessened and they will be able to foresee patients' illnesses with ease. The purpose of disease prediction is to make predictions for a variety of common illnesses that, if left untreated or neglected, can become deadly and create a great deal of trouble for the patient and their loved ones.

https://doi.org/10.5281/zenodo.14066308

The project's architecture allows the system to forecast disease by using the user's symptoms as input and producing output. In summary, the variety feature of hospital data determines how accurate risk prediction is for disease risk modelling.

## FUTURE SCOPE

- The ability to change user information.
- An interface that is more interactive.
- Tools for creating backups.
- It may be used as a webpage.
- It is possible to create a mobile application.
- Additional Information and Current Illnesses.

## REFERENCES

1. "M. Chen, S. Mao and Y. Liu. Big data: A survey".
2. "P. B. Jensen, L. J. Jensen and S. Brunak. Mining electronic health records: Towards better research applications and clinical care".
3. "Yulei wang1, Jun yang2, Viming.Big Health Application System based on Health Internet of Things and Big Data".
4. "S.-M. Chu,W.-T. Shih,Y.-H. Yang, P.-C. Chen and Y.-H. Chu. Use of traditional Chinese medicine in patients with hyperlipidemia: A population-based study in Taiwan".
5. "S. Zhai, Chang, R. Zhang and Z. M. Zhang. Deepintent: Learning attentions for online advertising with recurrent neural networks".
6. "M. Chen, Y. Ma, J. Song, C. Lai, and B. Hu. Smart clothing: Connecting human with clouds and big data for sustainable health monitoring".
7. "W. Yin and H. Schutze. Convolutional neural network for paraphrase identification".
8. "Weixing Wang and Shuguang Wu. A Study on Lung Cancer Detection by Image Processing".
9. "Thanh Trung Giang, Thanh Phuong Nguyen and Dang Hung Tran. Stratifying Cancer Patients based on Multiple Kernel Learning and Dimensionality Reduction, 2017 IEEE 9th International Conference on Knowledge and Systems Engineering (KSE)".
10. "Saranya P and Satheeskumar B. A Survey on Feature Selection of Cancer Disease Using Data Mining Techniques", International Journal of Computer Science and Mobile Computing, Vol.5 Issue.5, May- 2016, pg. 713-719".
11. "Dmitry Ignatov and Andrey Ignatov. Decision Stream: Cultivating Deep Decision Trees", 3 Sep 2017 IEEE".
12. "Kelvin KF Tsoi1, Yong-Hong Kuo and Helen M. Meng. A Data Capturing Platform in the Cloud for Behavioral Analysis Among Smokers An Application Platform for Public Health Research", 2015 IEEE".