CURIS: The Disease Predictor

ARCHIT SHARMA
DEPT. OF ISE, EWIT

ARGHADEEP BANERJEE
DEPT. OF ISE, EWIT

KRISHNAM CHATURVEDI
DEPT. OF ISE, EWIT

Mrs. SHRUTHI T V
ASST. PROFESSOR
DEPT. OF ISE, EWIT

ABSTRACT-- **CURIS: The disease prediction system is based on predictive modeling which predicts the diseases of the user based on the symptoms. The disease prediction system has three users such as doctor, patient and admin. Each user is authenticated by the system, there is a role-based access to the system. The system allows the patient to give symptoms. On the basis of the given symptoms, the system analyses the symptoms runs multiple algorithms and then gives the probability of the occurrence of a disease as an output. Disease prediction is done by using Supervised Machine Learning, the classifier calculates the probability of the disease and adds the accuracy score. The system also suggests doctors for predicted disease, where the user can consult with the doctor at their convenience by sitting at home. The chat box functionality will help the patients to take early measures before an appointment and thus help in saving themselves from getting things worse. The most prominently discussed supervised ML algorithms were Naive Bayes (NB), Decision Trees (DT).**

Keywords—Supervised Machine Learning, Diseases Prediction, Naive Bayes (NB),Decision Trees (DT).

I. INTRODUCTION

The emergence of Artificial Intelligence (AI) enabled computerized systems to perceive, think and operate in an in-Telligent manner like humans. AI is a multidisciplinary concept of ML, Computer Vision, Deep Learning, and Natural Language Processing. ML algorithms apply various optimization, statistical, and probabilistic techniques to learn from data that was generated from past experiences, and deploy it in decision making. These algorithms deemed to be applied in many disciplines including network intrusion recognition, customer purchase behavior detection, process manufacturing optimization, credit card fraud detection, and disease modulation. Many of these applications have been designed using the supervised learning approach. In this approach, datasets with known labels are induced to prediction models to predict unlabeled examples This presents the hypothesis that medical doctors can utilize supervised learning as a powerful tool to conduct diseases diagnosis more efficiently. The healthcare industry produces large amounts of healthcare data daily that can be used to extract information for predicting disease that can happen to a patient in future while using the treatment history and health data. This hidden information in the healthcare data will be later used for affective decision making for patient’s health. Also, these areas need improvement by using the informative data in healthcare.

One such implementation of machine learning algorithms is in the field of healthcare. Medical facilities need to be advanced so that better decisions for patient diagnosis and treatment options can be made. Machine learning in healthcare aids the humans to process huge and complex medical datasets and then analyze them into clinical insights. This then can further be used by physicians in providing medical care. Hence machine learning when implemented in healthcare can leads to increased patient satisfaction. The main focus is on to use machine learning in healthcare to supplement patient care for better results. Machine learning has made easier to identify different diseases and diagnosis correctly. Predictive analysis with the help of efficient multiple machine learning algorithms helps to predict the disease more correctly and help treat patients.

II. EXISTING SYSTEM

The existing system proposes various features to be implemented in health care with various benefits. At first Machine learning algorithms are used to analyze medical data sets effectively in present. The main aim was to implement supervised machine learning concept by using datasets regarding blood cells collected from blood cells detecting and counting sensors, of a human as the input. This input was trained by artificial neural network algorithm and decision tree classification learning algorithm to perform classification It resulted in recognizing and also possibly predict the disease based on the nature of blood cells and classify accordingly. Later

The application of machine learning in the field of medical diagnosis is increasing gradually. This contributed primarily to the improvement in the classification and recognition systems used in disease diagnosis. Different classification algorithms are applied, on three separate databases of disease (Heart,Diabetes) available in UCI repository for disease prediction. The feature selection for each dataset was accomplished by backward modeling using the p-value test. ML is playing a vital role in health care.
Implementation of ML technologies in health care as a part of prediction model helps doctors and patients to analyze the disease. The sub diseases fall under the same category. So, it becomes difficult to differentiate between diseases of the same category but of different types. The data sets include the data of mostly the frequently occurring diseases so it leaves out rare diseases or less occurring diseases. No sorts of precautions available to patients in case they cannot immediately pay a visit to the doctors. Existing system reviews concepts of Machine learning (ML), ML prediction model and applications, describes implementation of ML applications in health care. The existing system proposes various features to be implemented in health care with various benefits. At first Machine learning algorithms are used to analyze medical data sets effectively in present. The main aim was to implement supervised machine learning concept by using datasets regarding blood cells collected from blood cells detecting and counting sensors, of a human as the input.

III. PROPOSED SYSTEM

In this project, we have combined the structure and unstructured data in healthcare fields that let us assess the risk of disease. The approach of the latent factor model for reconstructing the missing data in medical records which are collected from the hospital. And by using statistical knowledge, we could determine the major chronic diseases in a particular region and in particular community. To handle structured data, we consult hospital experts to know useful features. In the case of unstructured text data, we select the features automatically with the help of Decision Tree algorithm. We propose a Decision Tree algorithm for both structured and unstructured data.

3.1 The Decision Tree algorithm

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. The decisions or the test are performed on the basis of features of the given dataset.

It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions. It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure. In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm. A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees.

Decision tree uses tree structure and the tree begins with a single node representing the training samples. If the symptoms are all of the same disease, then node becomes the leaf and the class marks it. Otherwise, the algorithm chooses the discriminatory attribute as the current node of the decision tree. According to the value of the current decision node attribute, the training samples are divided into serval subsets, each of which forms a branch, and there are serval values that form serval branches. For each subset or branch obtained in the previous step, the previous steps are repeated, recursively forming a decision tree on each of the partitioned samples.

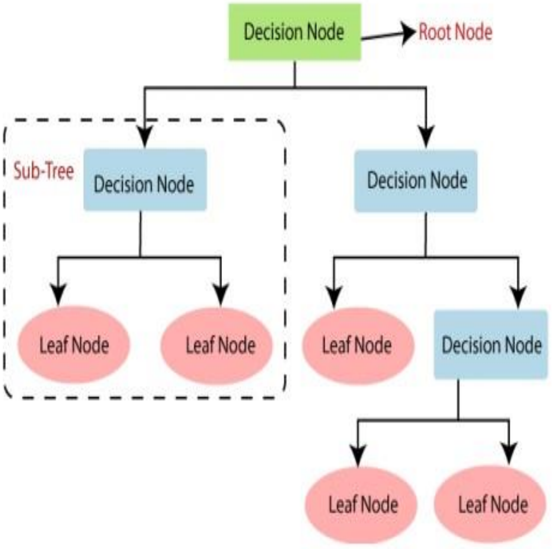


Fig.1. Decision Trees

All attributes are considered and the decision nodes take decision based on a criterion that estimates the cost. The best decision is the decision with the least cost or best accuracy.
The training data is also split based on these cost functions.
This algorithm can work recursively because at all decision node same steps need to be repeated. Cost functions are used for classification and regression. Either way, they strive to find branches with common criteria to decide the path this data frame belongs to.

Decision tree uses tree structure and the tree begins with a single node representing the training samples. If the symptoms are all of the same disease, then node becomes the leaf and the class marks it. Otherwise, the algorithm chooses the discriminatory attribute as the current node of the decision tree. According to the value of the current decision node attribute, the training samples are divided into several subsets, each of which forms a branch, and there are several values that form several branches. For each subset or branch obtained in the previous step, the previous steps are repeated, recursively forming a decision tree on each of the partitioned samples.

3.2 The Naïve Bayes algorithm

Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. It is mainly used in text classification that includes a high-dimensional training dataset. Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object. Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles.

The Naïve Bayes algorithm is comprised of two words Naïve and Bayes, which can be described as: Naïve: It is called Naïve because it assumes that the occurrence of a certain feature is independent of the occurrence of other features. Such as if the fruit is identified on the bases of colour, shape, and taste, then red, spherical, and sweet fruit is recognized as an apple. Hence each feature individually contributes to identify that it is an apple without depending on each other. Bayes: It is called Bayes because it depends on the principle of Bayes' Theorem.

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)}$$

Fig.2. Bayes Theorem

P(A|B) is Posterior probability: Probability of hypothesis A on the observed event B.

P(B|A) is Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true. Naïve Bayes is one of the fast and easy ML algorithms to predict a class of datasets. It can be used for Binary as well as Multi-class Classifications. It performs well in multi-class predictions as compared to the other Algorithms. It is the most popular choice for text classification problems.

Naïve Bayes classifiers are a collection of algorithms based on Bayes' theorem. The dataset is divided into two parts namely Feature matrix and Response vector. Feature matrix contains all vectors(rows) of the dataset in which each vector consists of value of dependent features. In Our data set, features are the symptoms such as 'Fever', 'Cold', 'Headache' etc. confusion matrix contains the values of Class variable (prediction or the output) for each row of feature matrix in our dataset, the class variables are the diseases like 'malaria', 'dengue' etc. Bayes' theorem will also find the probability of an event occurring and hence in our model it not only helps in predicting the disease but also gives the probability of the occurrence of the disease.

IV. SYSTEM ARCHITECTURE

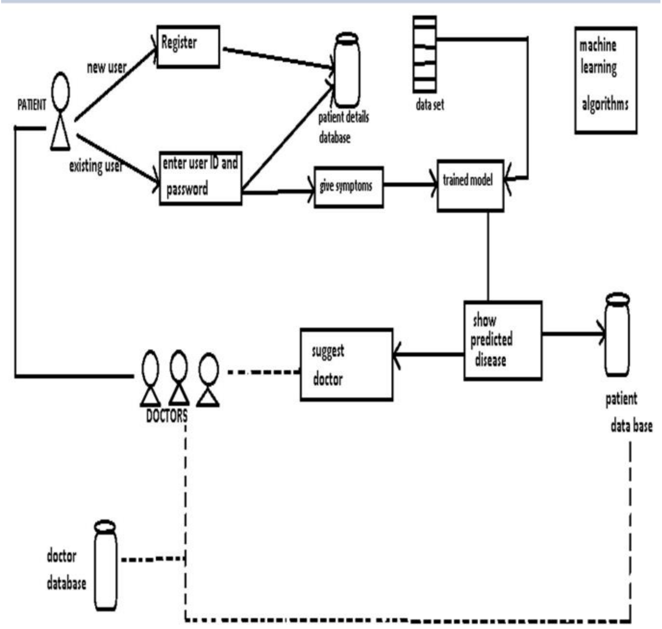


Fig.3. System Architecture

In architectural model it contains two databases: Patient Records database and Disease/Symptoms database. Four web services are used to implement the SOA. They are Pattern matching, recent trends, differential diagnosis and recent differential diagnosis. The patient Record database contains all the patient information from all the hospitals in the network.
Diseases/Symptoms database is a centralized database. It contains the list of existing known diseases and their corresponding symptoms along with their weights. These databases are replicated across various servers and these replicated servers are used to achieve the fault tolerance with concurrency protocols to achieve atomic transactions. First the doctor retrieves the symptoms from the patient record database. After retrieving the symptoms, the doctor identifies whether any symptom related diseases contain in the Diseases/Symptoms database. Here the pattern matching service is activated. If any diseases match with Symptoms

means list out all the possible matched symptoms and presents the result to the doctors. If the doctors not satisfied with results, compare to recent history and recent trend service must be activated. This service makes use of the Diseases/symptoms database and Patient Record database and the result obtained from pattern matching service to get results. After comparing the diseases to the recent history, cluster the shortlisted diseases.

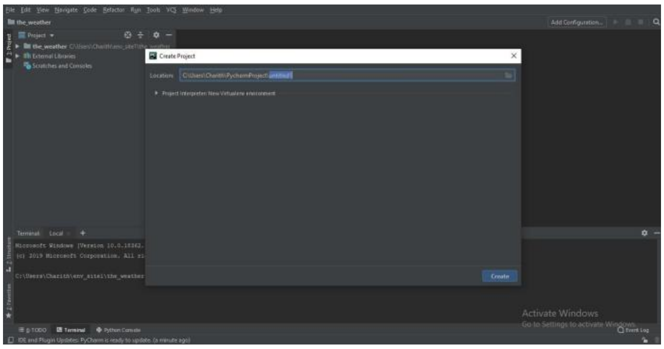
Finally, to avoid the vagueness in decisions, the doctor use differential diagnosis and recent diagnosis features use Diseases/symptoms database and Patient record database and result acquired from recent trend services to gain the results. Since the large medical data, using simple client server architecture would not produce the effective aforesaid services and would increase the response time of the system.

Finally, we conclude that SOA was well suited to apply this system because it improves the delivery of important information and sharing of data across the community of healthcare professionals more practical in cost, security and risk deployment. In various existing EHRs, SOA is more essential for data providers to this system, are already using this very successful and efficient architecture.

V. RESULT

The prediction model is developed for the disease prediction system using symptoms with chatbot and the accuracy is also good. The decision tree, Naïve Bayes algorithm is used for the better accuracy.

On the basis of the dataset of the symptoms we have and the input entered by the user the disease is predicted. The chat box helps as a interactive system for the user through which the diagnosis of the patient can be done earlier.

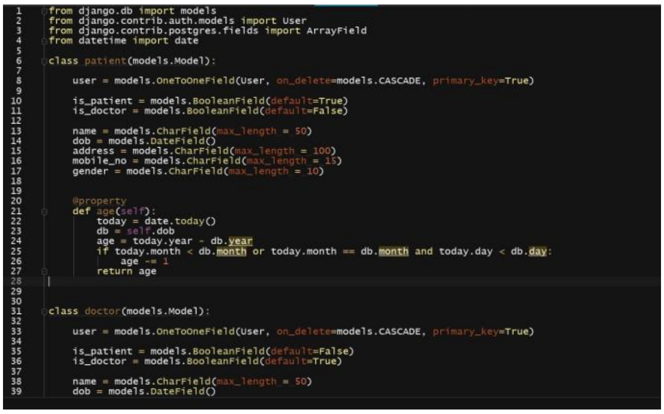

File Edit View Navigate Code Refactor Run Tools VCS Window Help
the_weather
Add Configuration...
Project
the_weather C:\Users\Charith\env_site1\the_weather
External Libraries
Scratches and Consoles
Create Project
Project Interpreter: New Virtualenv environment
Create
Terminal: Local
TODO
Terminal
Python Console
Event Log
Activate Windows
Go to Settings to activate Windows.

Location: C:\Users\Charith\PycharmProject\untitled

```
Microsoft Windows [Version 10.0.18362.
(c) 2019 Microsoft Corporation. All ri

C:\Users\Charith\env_site1\the_weather
```

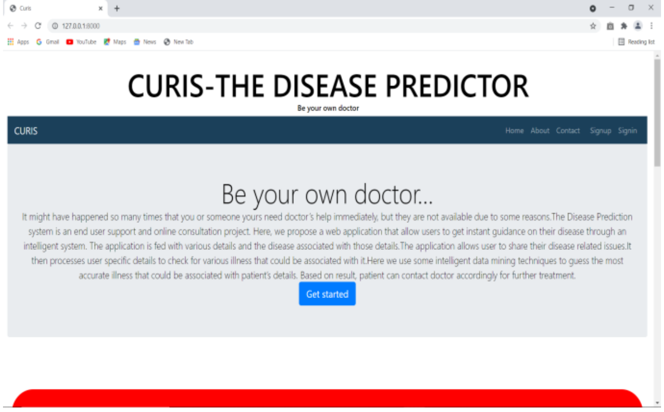
IDE and Plugin Updates: PyCharm is ready to update. (a minute ago)

Fig.4. Project Creation



```
from django.db import models
from django.contrib.auth.models import User
from django.contrib.postgres.fields import ArrayField
from datetime import date

class patient(models.Model):

    user = models.OneToOneField(User, on_delete=models.CASCADE, primary_key=True)

    is_patient = models.BooleanField(default=True)
    is_doctor = models.BooleanField(default=False)

    name = models.CharField(max_length = 50)
    dob = models.DateField()
    address = models.CharField(max_length = 100)
    mobile_no = models.CharField(max_length = 15)
    gender = models.CharField(max_length = 10)


    @property
    def age(self):
        today = date.today()
        db = self.dob
        age = today.year - db.year
        if today.month < db.month or today.month == db.month and today.day < db.day:
            age -= 1
        return age



class doctor(models.Model):

    user = models.OneToOneField(User, on_delete=models.CASCADE, primary_key=True)

    is_patient = models.BooleanField(default=False)
    is_doctor = models.BooleanField(default=True)

    name = models.CharField(max_length = 50)
    dob = models.DateField()
```

Fig.5. Connection to Database



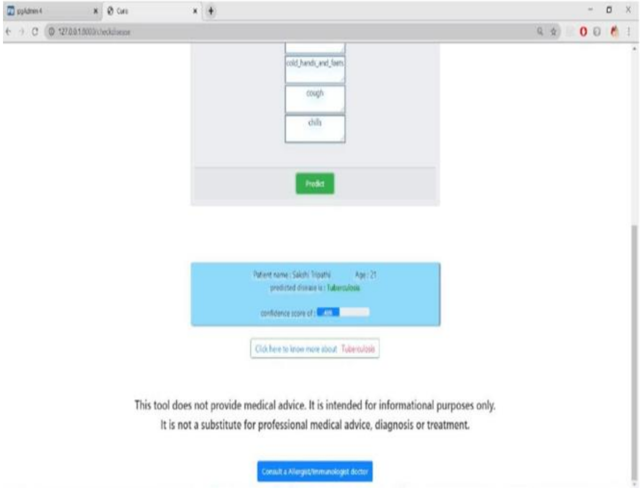
CURIS-THE DISEASE PREDICTOR

Be your own doctor

Be your own doctor...

It might have happened so many times that you or someone yours need doctor's help immediately, but they are not available due to some reasons.The Disease Prediction system is an end user support and online consultation project. Here, we propose a web application that allow users to get instant guidance on their disease through an intelligent system. The application is fed with various details and the disease associated with those details.The application allows user to share their disease related issues.It then processes user specific details to check for various illness that could be associated with it.Here we use some intelligent data mining techniques to guess the most accurate illness that could be associated with patient's details. Based on result, patient can contact doctor accordingly for further treatment.

Fig.6. Home Page


pgAdmin 4
Curis
127.0.0.1:8000/checkdisease
Predict
Consult a Allergist/Immunologist doctor

cold_hands_and_feets

cough

chills

Patient name : Sakshi Tripathi Age : 21

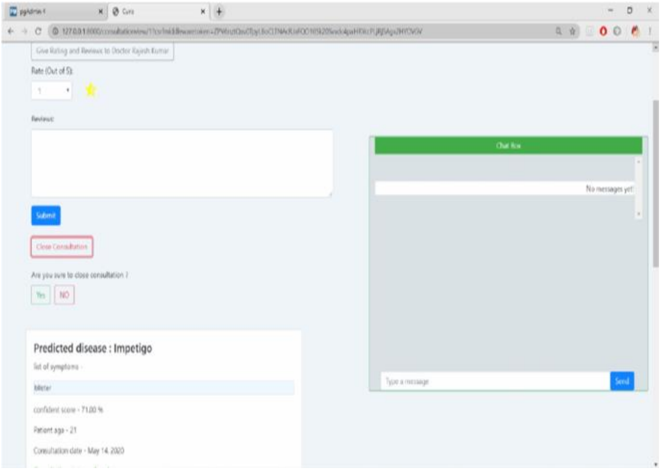
predicted disease is : Tuberculosis

confidence score of : 45%

Click here to know more about Tuberculosis

This tool does not provide medical advice. It is intended for informational purposes only. It is not a substitute for professional medical advice, diagnosis or treatment.

Fig.7. Symptom Adding Page



Predicted disease : Impetigo

list of symptoms -

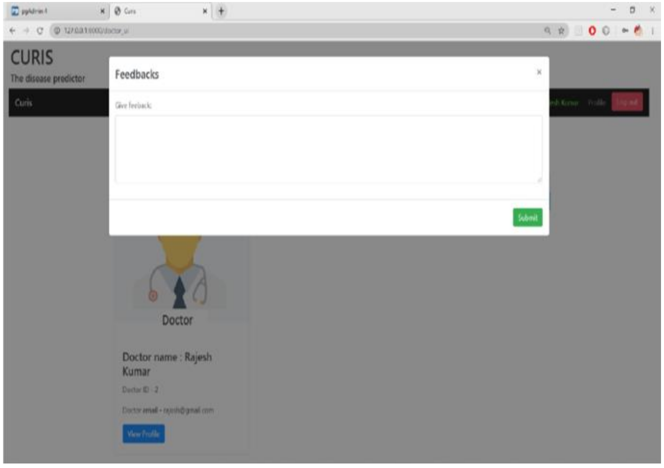
blister

confident score - 71.00 %

Patient age - 21

Consultation date - May 14 2020

Fig.8. Consultation page



Doctor name : Rajesh Kumar

Doctor ID - 2

Doctor email - rajesh@gmail.com

Fig.9. Feedback page

VI. CONCLUSION

Developing a prediction system using Machine learning, will be one of the reasons for technology to be closer to the patients and doctors, who can diagnose their disease by their symptoms and who can analyze the disease in better manner respectively. Health is important for human for achieving things greater in today's life.

Developing a prediction model using Machine learning also helps society by making people's life easier. Because at the end of the day no matter what a person learns in his lifetime, must be used in such a way that it helps him/her as well as the society. This prediction model can be a contribution to the society as well.

VII. REFERENCES

[1] D. W. Bates, S. Saria, L. Ohno-Machado, A. Shah, and G.Escobar, "Big data in health care: using analytics to identify and manage high-risk and high-cost patients," Health Affairs, vol. 33, no. 7, pp. 1123–1131, 2014.

[2] K.R.Lakshmi, Y.Nagesh and M.VeeraKrishna, Performance comparison of three data mining techniques for predicting kidney disease survivability International Journal of Advances in Engineering &Technology, Mar. 2014.

[3] Mr. Chala Beyene, Prof. Pooja Kamat, "Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Techniques", International Journal of Pure and Applied Mathematics, 2018.

[4] Boshra Brahmi, Mirsaeid Hosseini Shirvani, "Prediction and Diagnosis of Heart Disease by Data Mining Techniques", Journals of Multidisciplinary Engineering Science and Technology, vol.2, 2 February 2015, pp.164-168.

[5] A. Singh, G. Nadkarni, O. Gottesman, S. B. Ellis, E. P. Bottinger, and J. V. Guttag, "Incorporating temporal her data in predictive models for risk stratification of renal function deterioration," Journal of biomedical

[6] S. Patel and H. Patel, "Survey of data mining techniques used in healthcare domain," Int. J. of Inform. Sci. and Tech., Vol. 6, pp. 53-60,March 2016.

[7] F. Q. Yuan, "Critical issues of applying machine learning to condition monitoring for failure diagnosis," in 2016 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), 2016, pp. 1903–1907.

[8] S. Ismaeel, A. Miri, and D. Chourishi, "Using the extreme learning machine (elm) technique for heart disease diagnosis," in 2015 IEEE Canada International Humanitarian Technology Conference (IHTC2015), 2015,pp. 1–3.

[9] D. Dahiwade, G. Patle, and E. Meshram, "Designing disease prediction model using machine learning approach," Proceedings of the 3rd International Conference on Computing Methodologies and Communication, ICCMC 2019, no. Iccmc, pp. 1211–1215, 2019.

[10] S. Jadhav, R. Kasar, N. Lade, M. Patil, and S. Kolte, "Disease Prediction by Machine Learning from Healthcare Communities," International Journal of Scientific Research in Science and Technology, pp. 29–35, 2019.

[11] R. Saravanan and P. Sujatha, "A state of art techniques on machine learning algorithms: A perspective of supervised learning approaches in data classification," in 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), 2018, pp. 945–949.

[12] Y. Amirgaliyev, S. Shamiluulu, and A. Serek, "Analysis of chronic kidney disease dataset by applying machine learning methods," in 2018 IEEE 12th International Conference on Application of Information and Communication Technologies (AICT), 2018, pp. 1–4.